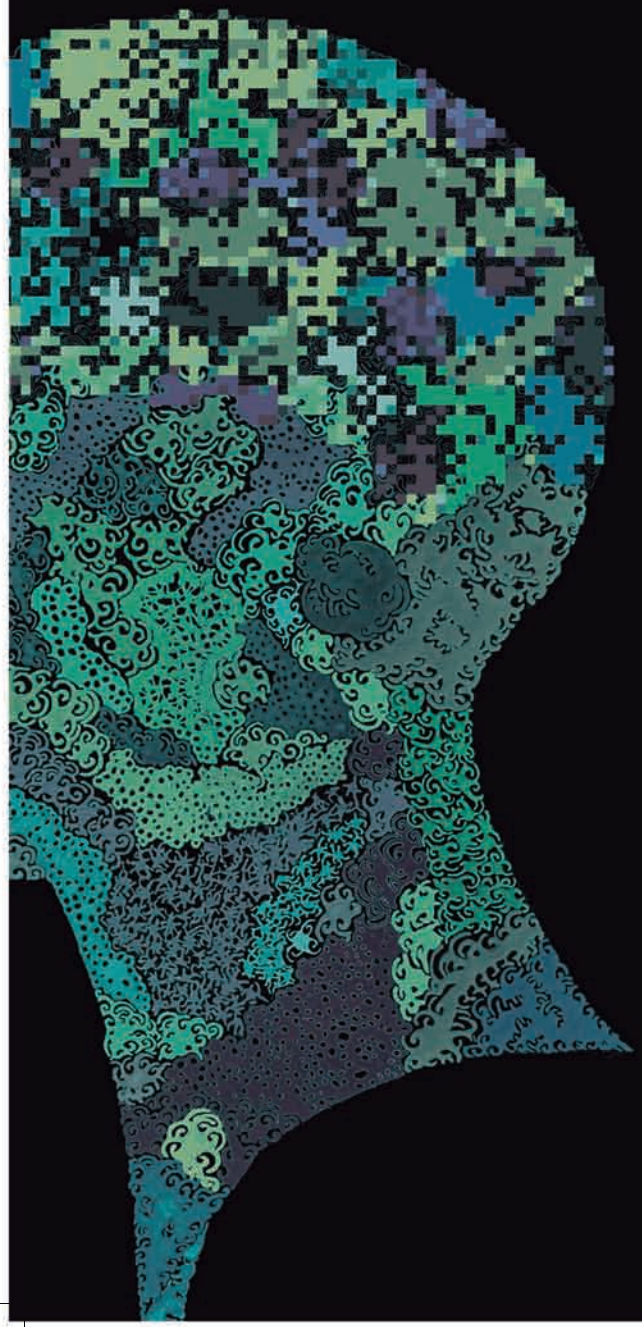


TESIS DOCTORAL Beatriz Serrano Solano AÑO 2017

TESIS DOCTORAL

MINERÍA DE DATOS APLICADA A LA MICROSCOPÍA DE SISTEMAS: ESTUDIO DE LA RELACIÓN ENTRE LAS ANOTACIONES FENOTÍPICAS Y FUNCIONALES EN CÉLULAS HUMANAS

Beatriz Serrano Solano

Director: Juan Antonio García Ranea
Programa de Doctorado: Biología Celular y Molecular
Facultad de Ciencias

UNIVERSIDAD DE MÁLAGA

TESIS DOCTORAL

Minería de datos aplicada a la Microscopía de Sistemas: Estudio de la relación entre las anotaciones fenotípicas y funcionales en células humanas

D.ª Beatriz Serrano Solano

Programa de Doctorado de Biología Celular y Molecular

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

Málaga, 2017

UNIVERSIDAD
DE MÁLAGA

AUTOR: Beatriz Serrano Solano

http://orcid.org/0000-0002-5862-6132

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

UNIVERSIDAD DE MÁLAGA

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

D. **Juan Antonio García Ranea**, Profesor Titular del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga,

CERTIFICA:

Que D.ª **Beatriz Serrano Solano**, Ingeniera en Informática por la Universidad de Málaga, ha realizado bajo mi dirección el trabajo de investigación correspondiente a su Tesis Doctoral titulada «**Minería de datos aplicada a la Microscopía de Sistemas: Estudio de la relación entre las anotaciones fenotípicas y funcionales en células humanas**».

Revisado el presente trabajo estimo que puede ser presentado al Tribunal que ha de juzgarlo.

Y para que conste a los efectos oportunos, firma el presente documento en Málaga, a 5 de Octubre de 2017.

Fdo: D. Juan Antonio García Ranea

UNIVERSIDAD DE MÁLAGA

Departamento de Biología Molecular y Bioquímica

Facultad de Ciencias

D. **Juan Antonio García Ranea**, Profesor Titular del Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga,

AUTORIZA:

A D.ª **Beatriz Serrano Solano** a la lectura y defensa de esta memoria de Tesis Doctoral. Asimismo, el director de la Tesis certifica que todas las publicaciones (y comunicaciones a congresos) que avalan la presente memoria de Tesis Doctoral no han sido utilizadas en tesis anteriores, y han sido generadas, parcial o totalmente, a partir de resultados y metodologías desarrollados durante la realización de la Tesis Doctoral de Dª. Beatriz Serrano Solano.

Y para que conste a los efectos oportunos, firma el presente documento en Málaga, a 5 de Octubre de 2017.

Fdo: D. Juan Antonio García Ranea

Este trabajo de Tesis Doctoral ha sido subvencionado por la red de Excelencia Europea «*Systems Microscopy*» y el proyecto AF2012-33110 del Ministerio de Economía y Competitividad (MINECO).

Parte de los resultados y metodologías recogidas en la presente Memoria han dado lugar a las siguientes publicaciones y comunicaciones:

- ✧ Serrano-Solano B., Díaz Ramos, A., Hériché, J.-K., & Ranea, J. A. G. (2017). How can functional annotations be derived from profiles of phenotypic annotations? BMC Bioinformatics, 18, 96. http://doi.org/10.1186/s12859-017-1503-5. (Véase anexo 7.2).

- ✧ Linking formal Ontologies & Experimental Phenotypic Profiles of Genes: A Comparative Study. NoE Systems Microscopy 4th Annual Meeting, Vienna (Austria), 2015. Poster.

- ✧ How loss-of-function phenotypes are related to gene functions. International Conference on Systems Biology 2016, Barcelona (Spain). Poster.

Agradecimientos

Decía *Erik Erikson* que soy lo que sobrevive de mí. Claramente, algunas partes de mí han muerto en este proceso, pero muchas otras se han desarrollado: pensamiento crítico, resiliencia y búsqueda de alternativas ante el fracaso. El cambio de disciplina tampoco ha sido un camino fácil, poco tienen que ver las ciencias de la computación con las ciencias de la vida. La ausencia de un conocimiento y vocabulario biológicos, junto con la necesidad de comprender esquemas de razonamiento distintos, me han proporcionado una perspectiva más amplia sobre la ciencia. En este proceso de transformación han tomado parte –directa o indirectamente– una multitud de personas a las que quiero agradecer particularmente su contribución.

En primer lugar a la Universidad de Málaga, por proporcionar el entorno de trabajo y las instalaciones para el desarrollo de esta tesis doctoral.

También a mi director, Juan Antonio, por darme la oportunidad de dedicarme a la Bioinformática en primer lugar. Y en segundo, por brindarme la posibilidad de desarrollar esta tesis dentro del marco de un proyecto europeo que me ha permitido viajar y relacionarme con investigadores de cuantiosa valía.

Jean-Karim Hériché has been a cornerstone of this thesis. He has helped me more than he could ever suspect. During my stay at EMBL, he taught me the values of honest and well-done science. But before and after that, while collaborating, I've learned so much from his solid knowledge and his abstract thinking, explaining unselfishly any doubts to anyone. Thanks, because you made this goal feasible.

A Antonio Díaz «el matemático», por estar siempre dispuesto a ayudar, a resolver cualquier duda y a mostrarme su apoyo y su amistad.

Dentro del departamento, aunque en disciplina ajena, Antonio Heredia siempre ha puesto en valor mis capacidades y me ha hecho entender la ventaja de llegar a una disciplina donde mi forma de pensar es completamente distinta, dándome ánimos para aprovechar ese potencial. Sus buenos consejos han sido siempre útiles y acertados. De su grupo salió Rocío, una impulsiva

y genuina mollinense que me acompañó en momentos de risa, pero también de angustia; a la que le deseo la mejor de las suertes y todo el ánimo para que pronto acabe su tesis en ese *possiblement nou país*.

Entre los compañeros del *«dark side»*, le debo un especial agradecimiento a la feminazi Rocío por darme muchas de las claves de esta tesis. Ella, junto con Aníbal, han sido los compañeros de remesa con los que he compartido momentos memorables de risas y algún lloro. En épocas distintas, Ian, Jim y Cristóbal me han apoyado y ayudado siempre que los he necesitado. Igualmente tengo mucho que agradecer a mis dos discutidores favoritos: David, amigo a pesar de nuestras continuas confrontaciones; y a Raúl, por sus comentarios directos, críticos y a veces tan hirientes como necesarios. Quiero extender asimismo mi agradecimiento a miembros anteriores: Almudena, Armando, Aurelio; y a los más recientes: Elena y Fernando. No me olvido tampoco de algunos miembros esporádicos que también me marcaron y en cierta manera hicieron esta experiencia más llevadera: Luz y Pedro, dos buenas personas que me crucé en el camino.

No a nivel científico pero sí de revisión y apoyo, el incansable y perfeccionista Marín, junto con la cándida Mili, han contribuido a la calidad de este manuscrito además de garantizarme siempre momentos de evasión.

Finalmente, a mi familia. Mis padres han vivido cada una de mis etapas con sufrimiento: desde una carrera compleja hasta una tesis que se ha presentado con no pocas dificultades. A mi hermano, por su capacidad innata de sobreponerse a todas las adversidades con buen humor. Y a David, por sus infinitas escuchas, por su serenidad, por ayudarme a relativizar y mostrarme que en mi mano siempre había una salida.

B.S.S.

Acrónimos

AP/MS	Affinity Purification/Mass Spectrometry
AUC	Area Under the Curve
BP	Biological Process
CC	Cellular Component
cDNA	complementary DNA
CMPO	Cellular Microscopy Phenotype Ontology
CPO	Cell Phenotype Ontology
CRISPR	Clustered Regularly Interspaced Short Palyndromic Repeats
DAG	Direct Acyclic Graph
dsRNA	double-stranded RiboNucleic Acid
EMBL	European Molecular Biology Laboratory
FDR	False Discovery Rate
GO	Gene Ontology
GOA	Gene Ontology Annotation
GMM	Gaussian Mixture Model
HCS	High-Content Screening
IC	Information Content
IDF	Inverse Document Frequency
IDR	Image Data Repository
IEA	Inferred from Electronic Annotation
IMP	Inferred from Mutant Phenotype
MAD	Median Absolute Deviation

MDS	MultiDimensional Scaling
MF	Molecular Function
MICA	Most Informative Common Ancestor
miRNA	micro RiboNucleic Acid
mRNA	messenger RiboNucleic Acid
OBO	Open Biomedical Ontologies
OWL	Ontology Web Language
PATO	Phenotypic Attribute and Trait Ontology
PCA	Principal Component Analysis
PCC	Pearson's Correlation Coefficient
PPI	Protein-Protein Interaction
RDF	Resource Description Framework
RISC	RNA-Induced Silencing Complex
ROC	Receiving Operating Characteristic
RNA	RiboNucleic Acid
RNAi	RiboNucleic Acid interference
siRNA	small interference RiboNucleic Acid
SOM	Self-Organizing Map
SS	Semantic Similarity
SVM	Support Vector Machine
TALEN	Transcription Activator-Like Effector Nuclease
TF-IDF	Term Frequency - Inverse Document Frequency
Y2H	Yeast Two-Hybrid

Summary

1. Introduction

One of the main goals of functional genomics is to unravel the functions of genes. The most reliable way to assign a function to a gene product is to perform a biological experiment. By inhibiting a gene, the resulting loss-of-function phenotype contributes to infer its biological function. The assumption behind this approach is that genes involved in the same biological process show similar loss-of-function phenotypes. This principle is used by the GO Consortium to annotate genes with terms from the biological process domain (i.e. annotations with evidence code IMP: Inferred from Mutant Phenotype).

Over the past decade, reverse genetics (sequence-specific genetic perturbations) has increasingly been performed, allowing large-scale targeted knock-down of genes. One of these techniques is RNA interference (RNAi) (Fire et al., 1998), a post-transcriptional gene perturbation mechanism. At a large scale, RNAi screens allow the systematic exploration of the effects of gene silencing and have become essential for assigning functions to genes due to their comparatively low cost and easiness compared to other techniques. Thousands of genes have been silenced in large RNAi screens (Neumann et al., 2010; Fuchs et al., 2010; Simpson et al., 2012).

In this context, Systems Microscopy approaches allow for systematic exploration of the gene loss-of-function phenotypic space, since it combines recent developments in microscopy automation with automated image analysis and data mining (Lock and Strömblad, 2010). These approaches differ from the traditional experiments: while typically the study focuses in a single phenotype tightly associated with a function, this novel approach uses phenotypic profiling to describe phenotypes by means of multi-parameter measurements. It yields much more profitable data since one single gene is linked to several measurements, yet the functional associations become less evident.

As a consequence, phenotypic annotations are usually manually converted into functional annotations. Given that the phenotypes are commonly based on free-text descriptions, the manual process is prone to errors and time-consuming, which leads to a low number of experimentally supported annotations with cellular functions. With

plenty of phenotypic annotations derived from high-throughput experiments, some of the gene functions have been revealed. However, this resource still remains highly under-exploited. RNAi screens in human cells are not commonly used to annotate genes in Gene Ontology with terms from the biological process domain. Thus, the number of experimentally-supported annotations is lower than the number of reported functional assays.

Apart from the drawbacks derived from the manual annotation process (e.g. time-consuming, error-prone, etc.), a second challenge is to assess the phenotypic similarity, particularly when integrating data from several experiments in which different variables are measured. Unlike the sequences or structures of gene products where the alignment algorithms provide a straightforward measure, finding a similarity value between two phenotypes that allows us to infer the same function is not always obvious.

The difficulties to functionally annotate each single gene product from high-throughput experiments make computational methods decisive for this task. Within this framework, several studies focused on the functional prediction indicates that genes predicted to be functionally related also show similar phenotypes (Lee et al., 2008; Qi et al., 2008; Hu et al., 2009; Rojas et al., 2012; Hériché et al., 2014).

Making data computationally tractable is a major requirement for automation. To deal with the data organisation and standardisation, ontologies are a key part. At the cellular level, the recently developed *Cellular Phenotype Ontology* (CPO, Hoehndorf et al. (2012)) and *Cellular Microscopy Phenotype Ontology* (CMPO, Jupp et al. (2016)) organise cellular phenotypes into a structured vocabulary. Ontologies are a way to represent knowledge that facilitates the automatic treatment of data, allowing reasoning and data exchange between different resources. Having a standardised set of phenotypic terms has an advantage over free-text phenotypic descriptions as it facilitates the automatic evaluation of phenotypic similarity.

On the basis of the abundance and heterogeneity of the experiments, if the conversion from phenotypic to functional annotations becomes automatic, then the wealth of data from the experiments could improve the understanding of biological and clinic knowledge.

2. Hypothesis and Goals

The automatic conversion from phenotypic annotations (in form of phenotypic profiles) to functional annotations requires to understand how phenotypes and functions are related. As a rule, loss-of-function phenotypes are widely used to infer gene function, but converting phenotypic to functional annotations demands a careful interpretation of phenotypic descriptions and assessment of phenotypic similarity. Understanding how functions and phenotypes are related provides insight into the development of methods for the automatic

conversion of gene loss-of-function phenotypes to gene functional annotations.

Most screens report hit lists of genes that are usually involved in the same biological process when looking for enrichment, i.e., similar GO annotations. In this way, genes phenotypically similar seem to be related to similar functions. We expect the phenotypic similarity to be indicative of participation in similar cellular processes. We wondered how integrated phenotypic profiles from multiple screens could be exploited to automate and/or refine the process of functional annotation.

Making use of the ontological structures as well as the computer tools for the interpretation of biological data, this work establishes the following goals to verify the hypothesis:

❶ Comparison of several phenotypic similarity measures and selection of the most appropriate to represent the functional relationship between gene pairs described through their phenotypic profiles.

❷ Study of the relationship between functional and phenotypic similarities of gene pairs to observe how phenotypic annotations can be derived from functional annotations and vice versa.

❸ Exploration of an alternative organisation to the ontological structure that relates both the phenotypic and functional spaces.

[1]http://www.ebi.ac.uk/cmpo
[2]http://www.ebi.ac.uk/fg/sym

3. Methods

In this work, we make use of several large siRNA-based experiments performed in human cells. We used genome-wide screens: *CellMorph* (Fuchs et al., 2010), which is focused on the morphological changes after the inhibition of each human gene; *MitoCheck* (Neumann et al., 2010), which study genes that might be involved in the chromosome segregation in HeLa cells; *EMBL secretion,* that analyses the transport from ER to the plasma membrane of the *ts045G* protein secretion. From the GenomeRNAi (Schmidt et al., 2013), we have used data from two genome-wide experiments: *GR00053* (Paulsen et al., 2009) y *GR00290* (Balestra et al., 2013), that analysed genome stabilization by phosphorylation of the histone H2AX in *HeLa* cells and the regulation of centriole biogenesis, respectively. These large screens were complemented by screens performed over specific subsets of genes to study the cellular response after ionising radiation in HeLa and U2OS cells (*Copenhagen DNA damage Ubiquitin*, (Moudry et al., 2012)) and the chromosome condensation (*EMBL chromosome condensation,* Hériché et al. (2014)).

The outcomes of these screens have been annotated with terms from the *Cellular Microscopy Phenotype Ontology* (CMPO) [1] (Jupp et al., 2016). The annotations were recorded in the *Cellular Phenotype Database*[2] (Kirsanova et al., 2015). Overall,

13866 links were obtained between 8109 genes and the following 36 unique phenotypes:

- ✧ abnormal chromosome segregation
- ✧ abnormal nucleus shape
- ✧ absence of mitotic chromosome decondensation
- ✧ binuclear cell
- ✧ bright nuclei
- ✧ cell death
- ✧ cell with projections
- ✧ decreased cell number
- ✧ decreased cell size
- ✧ decreased centriole replication
- ✧ decreased duration of mitotic prophase
- ✧ decreased number of site of double-strand break
- ✧ decreased rate of intracellular protein transport
- ✧ elongated cells
- ✧ graped micronucleus
- ✧ increased cell movement distance
- ✧ increased cell movement speed
- ✧ increased cell size
- ✧ increased cell size in population
- ✧ increased centriole replication
- ✧ increased duration of mitotic prophase
- ✧ increased nucleus size
- ✧ increased number of actin filament
- ✧ increased number of site of double-strand break
- ✧ increased rate of protein secretion
- ✧ increased variability of nuclear shape in population
- ✧ metaphase arrested
- ✧ metaphase delayed
- ✧ mild decrease in rate of protein secretion
- ✧ mitosis delayed
- ✧ mitotic metaphase plate congression
- ✧ more lamellipodia cells
- ✧ prometaphase delayed
- ✧ proliferating cells
- ✧ round cell
- ✧ strong decrease in rate of protein secretion

As the list of phenotypes illustrates, the cellular functions covered by these screens are wide: cellular proliferation, cell death, mitosis, protein secretion, DNA damage and centriole biogenesis. However, some other phenotypes not directly related to the process studied in the screen are also annotated, as in *MitoCheck*, where some phenotypes not directly related to the studied processes could appear.

It is noteworthy that the screens do not overlap in terms of phenotypes, which enriches the phenotypic profiles since the genes have been analysed from a wider variety of screens. All this data was integrated into a binary matrix with genes as rows and phenotypes as columns. Each cell of the matrix records the presence (value 1) or absence (value 0) of a phenotype for each gene. Given that not all screens are genome-wide, there are genes that have not been tested which are also represented in the matrix with value 0. This fact could affect our results, hence we tested the effect of sparsity by replacing a proportion (5, 10, 20 and 30 %) of randomly selected 1s by 0s.

To perform our study, we used a second formal ontology, *Gene Ontology* (GO) (Ashburner et al., 2000), from which we selected the branch of cellular processes, whose root term is *cellular process* (*GO:0009987*). The selection of this branch has been determined by the fact that the experiments

address cellular phenotypes. It is important to indicate that neither the genome-wide screens nor the screens performed over a specific subset of genes have been used to annotated genes in GO, as evidenced by the fact that none of the experiments articles is cited as a source of annotation in GO.

Genes were selected from the RNAi screens, considering that they were annotated in both GO and CMPO, i.e., only those with functional and phenotypic evidences have been included. In the *cellular process* GO branch we found 14080 annotated genes, of which 4198 are also annotated in CMPO. However, annotations in GO of genes with phenotypes may be biased. For instance, genes might be annotated with low informative GO terms or might be biased towards a particular branch due to the interest raised by disease genes or some widely studied biological processes. Thus, to ensure that the silenced genes showing phenotypes do not form a biased set of GO annotations, we compared two distributions of the information content (IC).

On one side, we calculated the IC distribution of the terms under the *cellular process* branch. This distribution showed how specific are the terms in which 14080 genes are annotated. On the other side, we computed the same IC distribution only for the functional terms under *cellular process* in which the 4198 studied genes are annotated (those with phenotypic annotations in CMPO).

The comparison of these two distributions revealed that both were very similar, but the one representing all the genes (14080) annotated in *cellular process* registered a higher density for low IC values (IC $\leqslant$ 2.5). A possible cause is a phenomenon called *"shallow annotation problem"*. This issue arises when genes whose function is unknown end up being annotated to very unspecific terms of the ontology. In short, the 4198 genes showing phenotypes are annotated in the ontology following a distribution similar to the original annotations in GO.

Once we verified that our set of genes was not biased, we measured the similarity between phenotypic profiles. This similarity is usually measured by applying vector-based measures such as *Euclidean, correlation* (Laufer et al., 2013; Bakal et al., 2007) or *cosine* (Loo et al., 2007; Wang et al., 2012). The *Euclidean* and *correlation* measures were calculated using the *R* package *stats*. For the *cosine* similarity the package *lsa2015* was used.

We also applied special similarity measures for binary characters such as *Hamming* and *Jaccard*. We used the *R* package *prabclus* (Hennig and Hausdorf, 2015) to calculate the *Jaccard* similarity and we implemented *Hamming* in *R*.

However, some phenotypes might be more informative than others within a phenotypic profile. This can be stated using similarities based on the information content. In this context, the information content refers to the specificity of a phenotype within a framework of annotations. The common basis to determine the specificity of a given phenotype relies on the low-frequency observation of this phenotype. For instance, *çell death"* (a widely observed phenotype), is considered less speci-

fic than *"mitotic delay"*, which is observed only in some specific cases.

Among the vector-based measures that consider the specificity of phenotypes we applied *Cohen's kappa* and *TF-IDF* (Robertson, 2004)), both implemented in *R*. In order to take advantage of the ontological structure of the phenotypes, we also calculated some semantic similarity measures (*Jiang*, *Lin*, *Schlicker* and *Resnik*) using the *R* package *dnet* (Fang and Gough, 2014).

All of the above mentioned measures were applied on the original matrix of genes and phenotypes. Nevertheless, when working in a highly dimensional space, it is sometimes beneficial to apply the vector-based similarities in a smaller dimensionality space. In our case, the phenotypic space contains 36 components or dimensions. Thus, we applied a dimensionality reduction method. Given that the matrix is binary (and hence the phenotypic profiles) the most appropriate method is the logistic PCA. After this, some of the similarity measures can be recalculated over the reduced space.

For dimensionality reduction of the matrix gene-phenotype, the *logisticPCA()* function of the *logisticPCA* package was applied, extracting the ten main components of the matrix. In this new space the *Euclidean*, *correlation* and *cosine* similarities were calculated.

Having this wide set of phenotypic similarities; we first evaluated the potential dependence between them. For this, we calculated the similarities between the values of each gene pairs for each measure. With these values, we computed the degree of relationship or similarity between the measures according to the *Pearson correlation coefficient* (*PCC*). Once the similarities between measures were obtained, they were organised into a hierarchical clustering structure using the *average-linkage* method of the *R* function *hclust*.

In addition to studying how measures relate to each other, these were also ranked by the ability to detect the functional relationship between genes. For this purpose, protein interaction networks were used as a proxy.

First, similarity measures were ranked using their ability to distinguish between interacting (true positive) and non-interacting (true negative) pairs by calculating the area under the ROC curve (AUC, Area Under the Curve).

As a positive set in the evaluation, we used the physical interactions of *Intact* (Orchard et al., 2014), *MIPS* (Pagel et al., 2005), *DIP* (Salwinski et al., 2004) and *BIOGRID* (Stark et al., 2006). From all resources, only those pairs identified by at least two different experimental methods were selected, or because they were curated interactions in *Reactome* ((Milacic et al., 2012; Fabregat et al., 2016)), resulting in 27689 relationships between genes in total.

As negative interaction set, we used the curated collection of interactions in the *MIPS Negatome* (Blohm et al., 2014) and Trabuco et al (Trabuco et al., 2012), which compose a set of 895790 nega-

tive relationships between gene pairs. To calculate the AUC, we used the *R* package *pROC*.

As a second approach to rank the similarities, we calculated the number of genes whose most phenotypically similar partner was also an interacting pair in *iRef index* (version 14.0, 04/04/2015) (Razick et al., 2008). The network of interactions consists of 22921 genes connected through 220926 relationships. In this approach, the method consists in identifying, for each measure and gene, the nearest neighbour (ties were broken at random). Only in the case that both genes were interacting partners in *iRef index* (in addition to being phenotypically close), will be considered to score the measure. In this way, the set of pairs with higher phenotypic similarities is expected to be enriched in physically interacting pairs.

A third alternative is the matrix comparison using the Mantel statistical test (Mantel, 1967). In this approach, the comparison between any phenotypic similarity matrix and the *iRef index* matrix was calculated using the *mantel* function of the *vegan* package.

With the appropriate measure between phenotypic profiles chosen, we studied how functional annotations were derived from the phenotypic profiles and vice versa.

Finally, following the alternative approach addressed by Glass and Girvan (Glass and Girvan, 2015), a bipartite graph was built with the GO functional terms and the phenotypes in CMPO, linked by the genes annotated in both domains. By projecting the bipartite graph, two possible networks were obtained: one graph composed of GO functional terms and a second one composed by CMPO phenotypical terms, where the link between two terms exists only in the case where they share at least one gene. The edge weight was given by the number of genes sharing both terms.

4. Results

Goal 1: Comparison of phenotypic similarity measures

Provided the different methods in which the phenotypic similarity can be measured, we wondered whether those measures would be orthogonal to each other. To answer this, we computed the correlation coefficient (PCC) between the phenotypic similarities, for all gene pairs and measures. Then, we represented them in a hierarchical clustering structure.

The resulting dendrogram showed that the similarity measures fell separated into two groups: first, those measures based on semantic similarity (*Jiang*, *Lin*, *Schlicker* and *Resnik*) and a second group formed by vector-based measures (*cosine*, *Euclidean*, *correlation*, *Jaccard* and *Hamming*), with *Cohen's kappa* occupying an intermediate position. This confirmed the idea that these groups of measures evaluate the phenotypic similarity quite differently.

The dendrogram assesses the relationship between phenotypic similarities. In order to know

which of the measures shows the highest correlation with the functional information, we tested them using protein interaction networks. One way to do it is to consider the interactions between proteins as positive cases of the functional relationship between genes. This means that when two proteins interact, the corresponding genes can be considered to be participating in the same function (Sharan et al., 2007). Thus, for a relevant measure, it is expected that phenotypically similar genes are enriched in interactions between proteins.

Three different tests were used here. The first one verifies the ability of each measure to distinguish between interacting and non-interacting pairs by calculating the area under the ROC curve (AUC). In a second approach, the nearest neighbour was identified –the most phenotypically similar gene– for each similarity measure and gene, and we checked whether those two genes were also interacting pairs in a protein interaction network. Finally, with the Mantel statistical test, we evaluated the correlation between the protein interactions matrix and each of the phenotypic similarity matrices.

After addressing these three approaches, *Resnik's* semantic similarity measure in CMPO ontology was the most consistent by associating similar phenotypes with interacting proteins, since it appears in all three cases at the top positions of the ranking of measures.

The rest of semantic similarities did not perform well. They may have been negatively influenced by the dispersion of the CMPO ontology since these measures take the topology into account more than *Resnik* do. *Lin* (Lin, 1998) and *Jiang* (Jiang and Conrath, 1997) are particularly sensitive to variations in the ontology, unlike *Resnik* (Resnik, 1995). This means that for a pair of terms, *Resnik* computes the IC of the most informative common ancestor (MICA), regardless of the distance to which that ancestor is from the pair of terms. However, *Lin* and *Jiang* take this distance into account, and therefore these two methods are more sensitive to the topological variations of the ontology. Since CMPO is an ontology with few levels, *Resnik* seems to be comparatively more robust.

Goal 2: Study of the relationship between functional and phenotypic similarities

Once the most appropriate metric to measure the similarity between phenotypic profiles was selected, we explored how the genes' functions related to the experiments in a more explicitly way. If phenotypes predict biological functions, it is expected that gene pairs with similar phenotypes share similar functions.

As the functions of the genes have been standardised using GO, the functional similarity between genes was computed using *Resnik's* semantic similarity between GO terms. This method seems to be the most appropriate to calculate the similarity in GO (Guzzi et al., 2012).

To evaluate the relationship between the phenotypic and functional similarities of genes in

GO, we plotted *Resnik's* semantic similarity in GO against *Resnik's* semantic similarity in CMPO for all gene pairs obtained in RNAi screens, excluding those genes without functional annotation in GO.

The distribution of the functional similarity values obtained remained constant for all levels of phenotypic similarity, except for high values corresponding to the range (6, 7.61], where a slight trend towards high values of functional similarity appeared. These pairs of phenotypically similar genes are also functionally related. However, we expected a clearer correlation between the functional and phenotypic similarities.

The weak trend that appears for high values of functional similarity could be due to chance. To verify it, we performed 1000 randomizations of the values with high similarity in CMPO (those with the phenotypic similarity that falls within the range (6, 7.61]). We computed the average semantic similarity for the original distribution and obtained 2.98. After the randomization of the values assigned to each pair of genes, a distribution with a mean of 1.53 was obtained. This indicates that the slight trend observed when evaluating the correlation between CMPO phenotypic similarity and GO functional similarity GO was not due to chance.

However, despite not being by chance, the number of genes responsible for this trend is low: there are only 20 unique genes for the range (6,7.61]. Therefore, taking into account that the overall set consists of 4198 genes, the proportion of genes with very specific phenotypes linked with specific functions is small. It seems that overall phenotypes are not a good predictor of biological function.

We also evaluated the effect of the evidence codes used when annotating the genes in GO. When a gene is linked to a term in GO by automated methods, the evidence code is IEA (*Inferred from Electronic Annotation*). The annotations with this non-curated evidence are usually discarded as they are generally considered less reliable, although in most cases, they have a positive or null effect on the results (Guzzi et al., 2012). In this particular case, we thought that these evidences could be influencing the absence of a trend, but when evaluating and including these IEA annotations, the trend did not significantly improve. Therefore, we can state that annotations with IEA evidence code do not affect the results obtained in a remarkable way.

Since the protein interactions have been considered as a proxy for the functional relationships between genes, it might be thought that the trend can become more pronounced when only keeping phenotypic similarity values for those pairs with a verified physical interaction. However, this did not happen, although in general the levels of semantic similarity in GO were slightly increased, the growing trend remained unchanged. Thus, the signal was still insufficient to establish a correlation between the two measures.

Up to this point, the analyses have been focused on the study of the relationships between the semantic similarity in GO –the most appropriate

way to measure the functional similarity between genes–, and the phenotypic similarity in CMPO. In parallel, we tested some of the above-mentioned phenotypic similarities between phenotypic profiles. Among all the phenotypic measures, the only one showing a weak trend was TF-IDF, while the rest of the measures did not present any kind of trend. In fact, the behaviour of TF-IDF is quite similar to the one shown by *Resnik's* phenotypic similarity in CMPO, because both works similarly: *Resnik* in CMPO selects the most specific common term and TF-IDF selects the common phenotype with the highest IDF (the most specific common phenotype).

One possible explanation for the lack of trend found in several tests is that some functions might be sharing the same phenotype. This would mean that one single phenotype is obtained when genes involved in different functions are silenced. As an example, this makes sense when knocking down a cell cycle gene and a gene involved in protein secretion, both resulting in the same phenotype (e.g. cell death).

If that were the case, similar functions should still lead to similar phenotypes. We would then expect that genes involved in the same cellular process would have similar phenotypes. When plotting the phenotypic similarity versus the functional similarity, we observed that it does not happen: genes with high functional similarity do not usually conduct to high phenotypic similarity.

So far, the results are not very promising. To reject the possibility that the high dimensionality of the matrix could be introducing noise, we applied PCA to the original gene-by-phenotype matrix. In this new space, we computed some vector-based similarities, but the results did not change either.

With the analysis performed so far –the selection of the most informative phenotypic term and the complete profile–, we have not found any indication of a significant relationship between phenotype and function. This result is contrary to intuition since most screens are based on the tenet that genes with the same biological function result in the same phenotypic description.

The experiments based on the phenotypic profiling states that each phenotype is indicative of an individual function. Therefore, the hypothesis formulated here is that single phenotypes are more informative than the phenotypic profiles.

To test this, we computed the average semantic similarity in GO for all gene pairs showing a given phenotype. Once this was done, this average was compared with the one obtained by randomising 100 times the associations between genes and phenotypes, while preserving the number of edges per phenotype.

A total of 8 out of 36 phenotypes (around 25%) gave a statistically significant signal (FDR-corrected p-value <= 0.01) for having their functional similarity between gene pairs above the one obtained by randomization. More than half of them correspond to CMPO terms with high information

content, indicating that only the most specific phenotypes tend to be associated with high similar functional annotations in GO.

The main conclusion we reached is that the automatic conversion of phenotypic annotations to functional annotations is only possible for highly specific phenotypes.

In order to discard that significant phenotypes were the only cause of the trend that appears to represent functional versus phenotypic similarity, the group of significant phenotypes was temporarily excluded from the gene-phenotype matrix. If these phenotypes significantly influence the trend, the weak tendency should fade after excluding them. However, in the results the outcome remains practically unchanged, indicating that the observed trend is not only an effect of these highly specific phenotypes, indicating that there is still information on function in the remaining phenotypes.

Goal 3: Exploration of an alternative organisation to the ontological structure

The results obtained so far suggest that the GO structure does not adequately reflect the functional relationships underlying the phenotypic similarity of RNAi screens. In this regard, an alternative way of organising GO terms has recently been proposed (Glass and Girvan, 2015).

The hypothesis is whether this approach would allow finding a relationship –stronger than that obtained so far– between functions and phenotypes. To test this hypothesis, we first grouped 7470 terms from the GO branch *cellular process*, in a total of 140 clusters. These 140 clusters represented groups of annotation-driven clusters of co-occurring GO terms, changing the definition of function to an integrated set of terms.

One way to evaluate whether this new definition of function relates to the phenotypic similarity is to measure *Resnik's* semantic similarity between the CMPO terms associated –through genes– to each cluster. As a null model, the associations between terms and clusters were randomised 1000 times. Regardless of functional clusters not linked to phenotypical terms, 77% (45/58) of the functional groups showed a high phenotypic similarity that can not be explained by chance.

Changing the approach by defining cellular functions as clusters of co-occurring GO terms allowed us to recover a stronger link between phenotypes and functions.

To verify the opposite assertion, i.e., whether similar phenotypes reflect similar functions, a graph of phenotypic terms was defined in an analogous way but in the CMPO ontology. This time, since the number of CMPO terms is 361, we define 13 clusters. The criterion here was to select a suitable number of clusters with more than one term.

The perspective is the opposite now: each phenotypic cluster can be seen as a phenotype characterized by a signature of co-occurring phenotypic descriptions. *Resnik's* semantic similarity between GO terms within clusters was calculated and again,

except for clusters that do not have GO annotations, it was observed that functional similarity was higher in phenotypic clusters than can be expected by chance. This demonstrates that this definition of phenotype allows recovering the functional similarity in GO.

Defining cellular phenotypes as groups of co-occurring CMPO terms also provided a link with functions. Thus, we can conclude that the cellular functions generated from the shared annotations were associated with the phenotypic similarity in CMPO.

5. Conclusions

In this work, we have explored how the phenotypic annotations of the genes from RNAi screens in human cells are related to the functional annotations in GO. After selecting the most appropriate measure to compare phenotypic profiles, we compared the similarities of gene pairs using GO and CMPO. We found that phenotypic similarity and functional similarity in GO do not correlate. However, by redefining functions as groups of co-occurring GO terms we were able to recover a stronger relationship between phenotypes and functions.

Therefore, several conclusions may be drawn from this Doctoral Thesis:

❶ The comparison of different similarity measures between phenotypic profiles determines that Resnik on the CMPO ontology is the measure that best represents the functional relationship between gene pairs.

❷ At a cellular level, it is not possible to derive the GO functional annotations of genes from the experimental phenotypic profiles or vice versa.

❸ Annotation-driven clustering of the ontological terms provides an alternative and complementary structure to the ontologies in which the phenotypic and functional spaces converge.

These results are particularly useful in data analysis and during the curation process after RNAi screens because the phenotypic similarity is used as a proxy for inferring gene function. For example, building annotation-driven clusters of phenotypes instead of grouping genes by phenotypic similarity (which have a low correlation with functional similarity) and then looking at the functional enrichment in those phenotypic clusters could be more informative.

Also, our results have implications when integrating the phenotypic data with other information resources for candidate gene selection. These methods are based on the combination of multiple sources to improve accuracy and coverage. Phenotypic information is often not used in these schemes but could be incorporated by using supervised machine learning methods. The training set might be defined by a two-inputs vector <*gene, phenotype*> and the expected output would be <*function*>.

The problem that arises here is that this association phenotype-function needs to be previously stated and has to be reliable because the result of the algorithm greatly depends on the quality of the training set. A proper design and quality of the training set are important when designing a kernel (e.g. for a kernel-based method, like SVM). Being consistent with the results presented here, using the usual vector-based measures for phenotypic similarity leads to poor performance in finding functionally related genes. According to our results, better phenotype kernels can be obtained if individual phenotypes are replaced by clusters of CMPO terms. Moreover, if diseases are seen as phenotypes, the functional similarity derived from the phenotypical clusters –guided by GO annotations– might be more useful for predicting disease-related genes than the usual functional similarity, which is based only on semantic similarity of individual CMPO terms.

Índice general

Índice de figuras

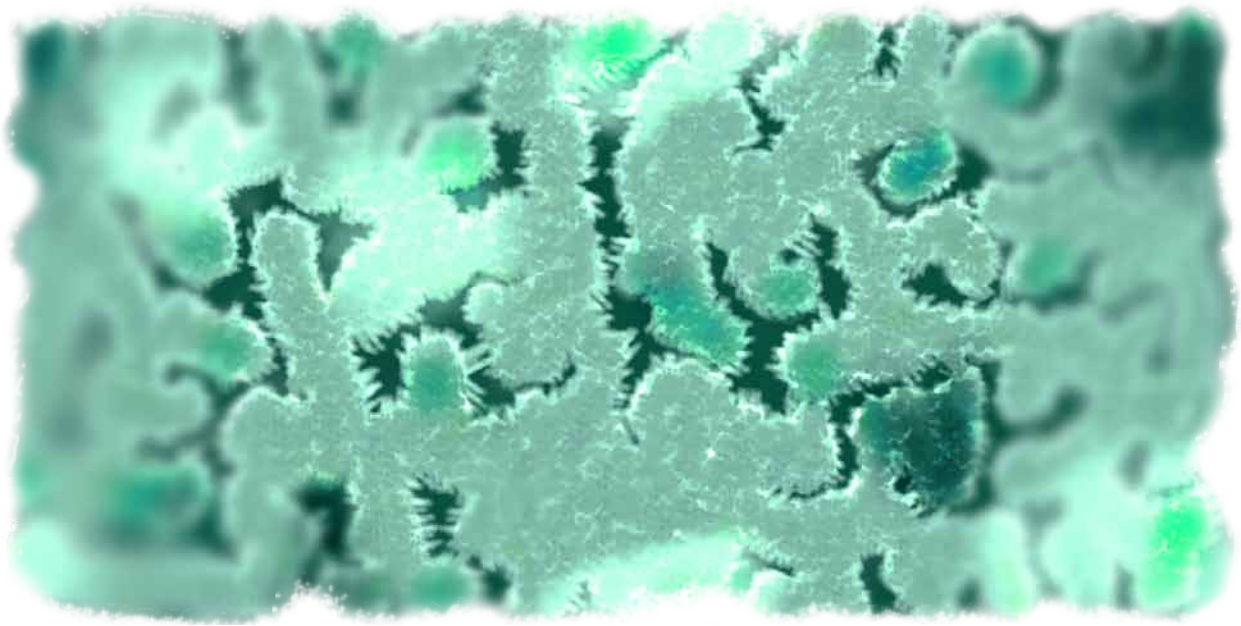
Índice de tablas

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

Isaac Asimov (1920 – 1992)

1

Introducción



En los últimos años ha surgido una revolución en el campo de la biología, provocada principalmente por la adquisición de una ingente cantidad de datos a través del desarrollo de nuevas tecnologías. Las medidas y observaciones recogidas a través de estas nuevas técnicas son, además de extensas, complejas. Surge, por tanto, la necesidad del procesamiento automático de datos que, de ser llevado a cabo de forma manual, sería un proceso laborioso y altamente propenso a errores. Este tratamiento, junto con el análisis y la interpretación de los resultados, desempeña un papel crucial en los avances de la investigación biomédica.

El fenómeno del *«big data»* ha contribuido de manera fundamental a la aplicación de técnicas informáticas en el campo de la biología, pero no es la única razón para su uso. Las necesidades de procesamiento que requiere la inteligencia artificial se han beneficiado de la paralelización, fruto del abaratamiento de la supercomputación. Esto, sumado a la madurez que han adquirido la inteligencia artificial y el aprendizaje automático o *«Machine Learning»* (Definición 1), ha facilitado el auge de la **bioinformática** o informática aplicada a diversos ámbitos biológicos.

> *Definición* 1. **Aprendizaje automático o *«Machine Learning»*:** rama de estudio de las Ciencias de la Computación que explora la construcción de algoritmos para el aprendizaje a partir de datos, permitiendo así generalizar y predecir datos futuros.

Las aplicaciones bioinformáticas a menudo se solapan con la Biología de Sistemas moderna –definida por (Kitano, 2002)–, que también avanza a pasos agigantados gracias al desarrollo de las técnicas *«ómicas»* (Definición 2). La capacidad de procesar grandes cantidades de datos hace posible abordar la complejidad de los sistemas biológicos, particularmente en el ámbito celular. Las técnicas *«ómicas»* proporcionan una valiosa información a distintos niveles, pero de forma independiente no tienen la capacidad de revelar completamente el funcionamiento de los sistemas celulares (Lock and Strömblad, 2010). Con una aproximación sistémica, se pueden combinar todas estas técnicas y considerar, por ejemplo, las variaciones inter-celulares, espaciales y dinámicas en el tiempo. Estos parámetros son fundamentales en la organización y el

comportamiento de un sistema celular (Lock and Strömblad, 2010). Por ello, se hace necesario integrar y complementar estas aproximaciones para cuantificar los procesos a escala celular.

> *Definición* 2. **Técnicas *«ómicas»*:** conjunto de disciplinas y áreas de investigación que estudian los sistemas biológicos a distintos niveles moleculares. Entre algunas de estas técnicas se encuentran la genómica, la proteómica, la transcriptómica, la metabolómica, la interactómica, la epigenómica y la metagenómica.

El desarrollo de nuevas técnicas de microscopía –junto con el análisis cuantitativo de las imágenes que se generan y la minería de estos datos– permite la detección, integración, análisis estadístico y modelado de eventos dinámicos que ocurren simultáneamente a diferentes niveles de resolución, desde la escala molecular hasta la celular. Este procedimiento y conjunto de técnicas es lo que se conoce como *«Microscopía de Sistemas»* (Lock and Strömblad (2010)).

La fase analítica es vital porque, como ocurre con las tecnologías *«ómicas»*, la *«Microscopía de Sistemas»* debe estar dirigida a la integración de datos y conocimiento procedentes de distintos experimentos, para lograr así un entendimiento en profundidad de los sistemas celulares. La principal ventaja que la *«Microscopía de Sistemas»* aporta sobre las técnicas *«ómicas»* es la posibilidad de integrar las dimensiones temporal y espacial en el análisis. Esta integración es especialmente útil para poner de relevancia fenómenos emergentes de auto-organización que aparecen como resultado de la interacción de moléculas dentro de las células (Lock and Strömblad, 2010).

Los estudios más habituales enfocados en la *«Microscopía de Sistemas»* suelen abordar el experimento en dos ámbitos distintos. Por un lado, están aquellos enfocados en el análisis de un sistema sub-celular específico; y por otro, los ensayos biológicos de alto rendimiento que introducen perturbaciones de forma discreta para identificar actores moleculares clave y su impacto funcional dentro de los procesos celulares complejos.

La visión más global la aportan los ensayos de alto rendimiento que, por lo general, aplican el mecanismo de inhibición mediante siRNA (*«small interference RNA»*). Esta técnica ha supuesto en los últimos años un avance en las aproximaciones para decodificar las funciones de los genes, con numerosas aplicaciones a nivel terapéutico (Hannon and Rossi, 2004). Sin embargo, esta

aproximación, aunque más completa, puede dar lugar a asociaciones difusas entre la función de un gen y el fenotipo observado tras su inhibición.

Dado que la cantidad de datos que se genera en estos experimentos es alta –y posiblemente contenga ruido–, su análisis ha de ser sistemático, para lo cual es necesaria la organización y clasificación de dichos datos. Para cubrir esta necesidad son útiles las ontologías (Definición 3), cuya función es definir un esquema de relaciones conceptuales a través de un vocabulario estructurado que favorezca la comunicación y el intercambio de información entre distintas fuentes. El uso que se le da actualmente a una ontología permite que se acceda a dichas entidades de forma computacional. Semánticamente, se trata de una forma de representar el conocimiento que facilita el tratamiento automático de los datos.

> *Definición* 3. **Ontología:** Descripción de un dominio de la realidad (biológica, por ejemplo) a través de un lenguaje estandarizado para clasificar y relacionar entidades.

La organización de la información proporciona numerosas ventajas. Por un lado, ayuda a clasificar los eventos celulares para facilitar su estudio. Y por otro, permite determinar la cercanía o lejanía entre los conceptos que clasifican un conjunto de genes. Esto último es particularmente útil para agrupar los resultados procedentes de experimentos de silenciamiento usando siRNA en función de los procesos en que un grupo de genes está involucrado. La agrupación semántica permite inferir relaciones desconocidas entre genes y plantear nuevas hipótesis a partir de resultados computacionales que arrojen luz sobre un determinado sistema celular.

1.1. Estado del arte

La secuenciación del genoma completo de muchos organismos ha propiciado la identificación de genes pero no necesariamente su caracterización funcional (Hancock (2014), Chapter 1). Asignar una función a dichos genes sigue siendo aún un reto, aunque las aproximaciones experimentales más recientes ponen el foco en encontrar la relación entre las características genómicas y los fenotipos. A esta disciplina se le llama *«fenómica»* (Hancock (2014)).

> *Definición* 4. **Fenómica:** área de estudio que describe los fenotipos de un organismo bajo la influencia de factores genéticos y ambientales.

Gracias al desarrollo de la secuenciación de alto rendimiento, de la anotación del genoma mediante técnicas computacionales y de la proteómica, la mayoría de los componentes básicos de la célula se han identificado (Hancock (2014), Chapter 5). Pero esto no es suficiente, es fundamental analizar cómo estos componentes interactúan para entender los procesos celulares.

El método tradicional para la caracterización de fenotipos celulares se basa en los experimentos de perturbación genéticos (Campbell and Bennett, 2016). Estas perturbaciones pueden venir dadas por distintas técnicas (Liberali et al., 2014) dependiendo de si actúan sobre el DNA (como el *knockout* de genes por ejemplo), sobre el RNA (a nivel post-transcripcional, como es el caso de los experimentos de RNAi) o bien a nivel de proteína (en la etapa post-traduccional, como en el *compound screening*). Cualquiera de estos métodos pueden aplicarse para establecer relaciones causales entre genes y fenotipos, además de proporcionar un método fiable para asignar funciones a genes. Por tanto, el fenotipado de células permite comprender las funciones celulares.

Para abaratar costes y abordar el problema de forma automatizada se aplican dos formas distintas de fenotipado sistemático que permiten ampliar el conocimiento del funcionamiento celular: una consiste en identificar todos los genes que contribuyen a un proceso biológico; mientras que otra posibilidad se basa en estudiar simultáneamente un rango de fenotipos.

En esta sección se describirán los últimos avances en el ámbito de desarrollo de esta Tesis Doctoral, abordando el proceso completo desde el silenciamiento génico para la observación de los fenotipos hasta su vinculación con la función que dichos genes desempeñan a nivel celular.

1.1.1. Definición de fenotipo

El término fenotipo (Definición 5) puede usarse en diversos ámbitos y a distintos niveles, abarcando desde un único rasgo hasta el conjunto completo de características de un organismo. Por ejemplo, en un contexto médico, al referirse a un fenotipo patológico se describe una característica distinguible de una población considerada normal. En el otro extremo, puede considerarse un fenotipo al perfil de expresión de una célula o a la concentración de un determinado metabolito en un medio.

> *Definición* 5. **Fenotipo:** conjunto de características observables de un organismo, comprendiendo su morfología, fisiología a nivel de célula, órgano u organismo, y su comportamiento (Nachtomy et al., 2007).

Aunque los fenotipos celulares se pueden definir a niveles muy distintos, desde la invención del microscopio se vienen describiendo mediante microscopía óptica. El desarrollo de técnicas para marcar moléculas dentro de la célula ha facilitado la obtención de dichas imágenes a escala celular. Aquí nos centraremos en los fenotipos obtenidos mediante microscopía óptica que se producen en la Biología Celular como resultado del silenciamiento génico.

1.1.2. Silenciamiento génico

El descubrimiento del RNA de interferencia (RNAi, *«RNA interference»*) podría catalogarse como uno de los eventos más importantes de los últimos años en el campo de la biología (Fire et al., 1998). La capacidad de silenciar un gen ha revolucionado la habilidad para decodificar la función de los genes, con las implicaciones que ello conlleva tanto a nivel terapéutico como para el desarrollo de la investigación básica.

El objetivo de los experimentos de inhibición de genes es la obtención de fenotipos que puedan ser ilustrativos de su función. Antes de la aparición de la técnica de inhibición RNAi, los experimentos de fenotipado se basaban en la sobrexpresión de proteínas usando cDNA (*«complementary DNA»*) (Okayama and Berg, 1982). Sin embargo, una de las limitaciones de esta técnica es que la sobrexpresión de una proteína no tiene por qué ser indicativa de una disfunción. Por ejemplo, podría ser que la función habitual del gen produzca un aumento de expresión, por lo que no sería distinguible de una ganancia de función.

Por estos motivos, por su sencillez de aplicación y por su bajo coste, el RNAi se ha convertido en la vía más común para producir fenotipos derivados de la pérdida de función (Liberali et al., 2014). Otras técnicas alternativas para el silenciamiento génico son TALEN (*«Transcription Activator-Like Effector Nuclease»*) (Gaj, 2014) y CRISPR (*«Clustered Regularly Interspaced Short Palyndromic Repeats»*) (Sternberg and Doudna, 2015). Esta última es tremendamente versátil y podría terminar imponiéndose en el futuro (Shalem et al., 2015), aunque actualmente coexisten y la utilización de una u otra viene dada por las necesidades del estudio (Boettcher and McManus, 2015).

El silenciamiento de la expresión génica mediante siRNA actúa de manera específica sobre una secuencia concreta cuya longitud oscila entre los 21 y 23 pares de bases (Figura 1.1). El mecanismo se inicia con la transfección de una doble cadena de RNA exógena (dsRNA o *«double-stranded RNA»*) en células eucariotas que silencia la expresión de un determinado gen, evitando que los mecanismos antivirales de la célula se activen, lo que induciría la célula a la apoptosis. Una vez introducido el dsRNA en la célula (Figura 1.1a), esta molécula es troceada por una enzima *Dicer* (Figura 1.1b), convirtiéndolo así en un siRNA funcional (Figura 1.1c), que se incorporará a un complejo de silenciamiento llamado RISC (*«RNA-Induced Silencing Complex»*)

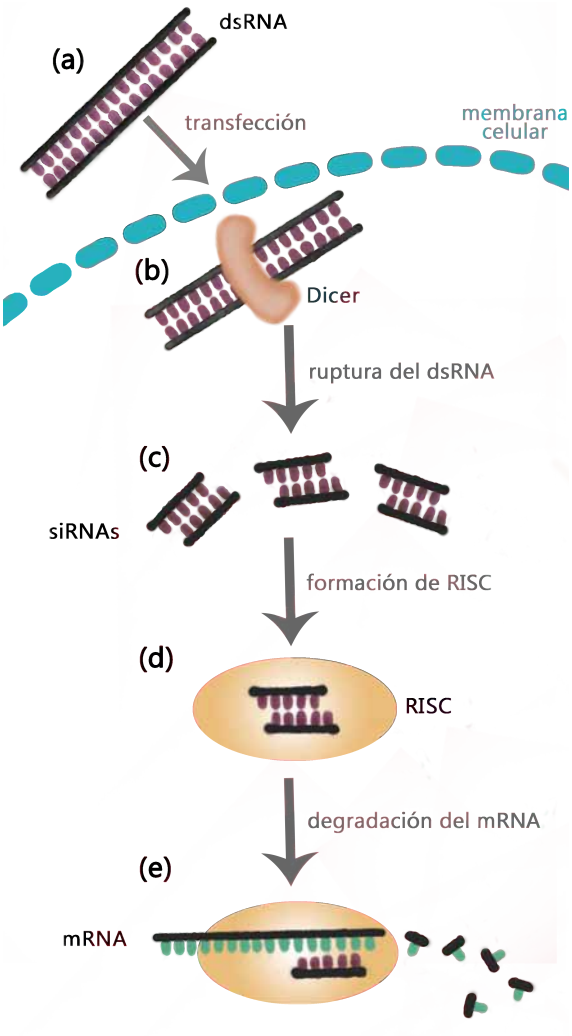


Figura 1.1: **Silenciamiento génico usando siRNA.** Partiendo de un dsRNA exógeno **(a)** que se introduce en la célula mediante transfección, la enzima *Dicer* digiere dicho dsRNA **(b)** y lo fragmenta en siRNAs funcionales **(c)**. Este siRNA funcional se incorpora al complejo RISC **(d)** que silencia al RNA mensajero **(e)**.

(Figura 1.1d), encargado de identificar y silenciar un RNA mensajero *(mRNA, «messenger RNA»)* complementario, causando su degradación y/o la inhibición de la traducción (Figura 1.1e). Es, por tanto, un mecanismo de silenciamiento génico post-transcripcional.

Uno de los principales problemas que pueden surgir es que la especificidad del silenciamiento no sea completa. Al igual que en el caso del micro RNA (miRNA, *«micro RNA»*), las dianas son reconocidas con un cierto valor de tolerancia en la complementariedad. Esto puede dar lugar a una inhibición no deseada de un RNA mensajero. Durante la síntesis de proteínas también podrían darse estas imprecisiones, que pueden deberse a la calidad de los reactivos, la muestra elegida, o incluso a una respuesta global de la célula. Esto es lo que se conoce co-

mo efectos *«off-target»* (Definición 6; Jackson and Linsley (2010)) y pueden reducirse usando distintos controles para minimizar los falsos positivos (Echeverri et al., 2006). Otra posible solución consiste en aplicar distintos siRNAs que inhiban el mismo gen. Así, la probabilidad de que varios siRNAs con secuencias completamente distintas –pero diseñados para inhibir al mismo mensajero– compartan los mismos efectos *off-target* es muy baja. Por todos estos motivos, es fundamental partir de un buen diseño experimental (Boutros and Ahringer, 2008).

> *Definición* 6. **Efecto *«off-target»***: conjunto de consecuencias fenotípicas detectables procedentes de interacciones no deseadas, ya sean dependientes de la secuencia de nucleótidos o no, entre las moléculas silenciadoras y diversos componentes celulares.

No obstante, las ventajas que esta técnica ofrece –principalmente derivadas de la facilidad de su aplicación– son mucho mayores que el grado de imprecisión asociado.

1.1.3. Anotación automática de fenotipos a partir de imágenes

Aunque los fenómenos que pueden darse en un sistema –desequilibrio de concentraciones de enzimas, marcadores, anotación de malformaciones detectadas por un médico, etc.– pueden registrase mediante su inspección directa, en las ciencias de la vida se usa a menudo la captura de imágenes. Así, se pueden determinar tanto la composición estructural y molecular, como la dinámica de las células, tejidos y organismos.

Los fenotipos detectables visualmente son muy útiles porque pueden ayudar a determinar la especificidad del efecto que produce un gen. Este proceso se ha automatizado en los últimos años gracias a los avances tecnológicos y se ha usado para observar el efecto de grandes bibliotecas de compuestos con potencial uso terapéutico (Li, 2003; Wilson, 2005; Granas, 2006; Gururaja et al., 2006; Richards et al., 2006), o para identificar y caracterizar la función de los genes (Lee et al., 2008; Qi et al., 2008; Hu et al., 2009; Rojas et al., 2012; Moreau and Tranchevent, 2012; Hériché et al., 2014). A este tipo de experimento se le denomina *High-Content Screening (HCS)* (Definición 7).

> *Definición* 7. **High-Content Screening (HCS)**: técnicas que combinan la microscopía automática junto con el posterior análisis de las imágenes adquiridas para el estudio de células vivas.

El resultado de estos experimentos no sólo es un conjunto de imágenes con un tamaño que varía desde megabytes hasta terabytes, sino que también se almacena el procedimiento que se ha llevado a cabo para la obtención de dichos datos (esto es lo que se conoce como *metadatos*, Definición 8). Por ejemplo, la placa y el pocillo donde se ha llevado a cabo el silenciamiento serían metadatos que identificarían un gen y el siRNA que lo ha inhibido.

> *Definición* 8. **Metadato**: información que se provee sobre otros datos con el objetivo de describirlos, organizar su estructura o ayudar a gestionarlos.

Una base de datos que almacena información fenotípica derivada de experimentos de silenciamiento génico es *GenomeRNAi* (Schmidt et al., 2013): en su versión del año 2013 contiene 170 experimentos en *Drosophila* y 127 en *Homo sapiens*. Además de los fenotipos, incluye me-

tadatos sobre los reactivos, junto con la eficiencia y especificidad como medidas de la calidad del experimento.

Una vez realizados los experimentos y almacenados los resultados, es necesario analizar de forma sistemática las imágenes obtenidas para extraer información sobre el proceso biológico en estudio.

El procedimiento seguido habitualmente consiste en anotar manualmente el fenómeno observado en la imagen. Esta anotación suele considerarse un criterio de referencia o *gold standard*. Sin embargo, esto presenta una alta propensión a errores. Uno de ellos es que la anotación varía dependiendo de la persona que observe el fenotipo. Para minimizar esta variabilidad, en algunos experimentos, una única persona da uniformidad a los resultados finales, aunque varias personas hayan anotado las imágenes previamente (Sönnichsen et al., 2005). No obstante, también pueden aparecer inconsistencias cuando la anotación la hace la misma persona en distintos días (Zhong et al., 2012). Otro sesgo se debe al conocimiento previo de la persona que anota. Por ejemplo, si se inhibe un gen que es conocido por estar involucrado en un proceso celular concreto, es posible que aunque el fenotipo no aparezca de una forma extendida en toda la muestra, la imagen sea anotada contando con la presencia del fenotipo esperado.

A la hora de detectar cambios cuantitativos en una imagen surge otra dificultad, ya que el ojo humano tiene limitaciones para percibir pequeñas variaciones en la iluminación, color, etc. Otro efecto bien conocido es debido a la saturación por el hecho de recibir información secuencial a través de la vista en forma de imágenes de imágenes (Miller, 1956). El canal visual posee una capacidad finita y muy limitada de procesar información de forma acertada, por lo que la anotación de las imágenes previas tiene influencia sobre la anotación en curso.

Aunque parece haber razones suficientes para el procesamiento automático de imágenes, el principal motivo para la automatización es la restricción temporal, ya que la alta cantidad de imágenes que se genera hace que no sea viable analizarlas de forma manual. Por ello, los algoritmos de visión artificial para el análisis de imágenes biológicas desempeñan un papel fundamental en el diseño de un proceso completamente automático que permita interpretar los experimentos biológicos más rápidamente y desde una perspectiva más objetiva.

1.1.3.1. Análisis de imagen

El método tradicional de análisis de imágenes distingue varias etapas que constan principalmente de la segmentación y la extracción de características, para finalmente clasificar e interpretar las imágenes en función de su contenido.

Durante la etapa de segmentación se establece un valor umbral para una función dada (detección de bordes, regiones, texturas, etc.), que permite subdividir la imagen en dos componentes homogéneos, esto es, separa el elemento de interés del fondo (Figura 1.2). Este valor umbral no tiene por qué aplicarse a toda la imagen por igual (global), sino que su valor puede variar por zonas de la imagen (dinámico) o depender de los píxeles vecinos (local).

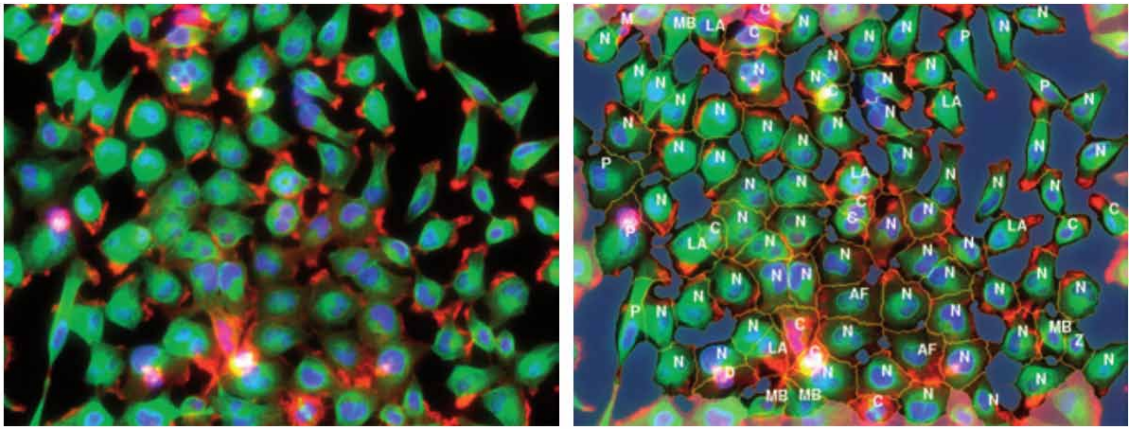


Figura 1.2: **Ejemplo de clasificación de células usando segmentación.** Células humanas sin procesar (izquierda) e imágenes tras la segmentación automática (derecha, donde cada letra corresponde a un tipo celular). Imágenes extraídas de Boutros and Ahringer (2008).

Una segunda etapa se centra en la extracción de características, en la que se consideran valores de un conjunto de funciones que caracterizan la imagen: iluminación, bordes, formas, texturas, etc.

La extracción de características abarca un concepto más amplio que la segmentación, ya que no se centra en la búsqueda de un objeto de interés, sino que proporciona un conjunto de descriptores que caracterizan la imagen. La segmentación es por ello dependiente del experimento: por ejemplo, para un ensayo donde se analiza un proceso que tiene lugar en el núcleo, la segmentación irá dirigida a esa parte de la célula, obviando el resto. En cambio, al extraer las características de la imagen completa, se obtiene un conjunto de valores que facilita la detección

de fenómenos que no se habían contemplado en el diseño experimental. Además, es importante destacar que los experimentos de alto rendimiento miden generalmente la señal promedio de todas las células dentro de un pocillo, por lo que no se tienen en cuenta las diferencias que podrían existir en las respuestas de células a nivel individual (Boutros et al., 2015).

La última etapa es la clasificación de las imágenes, para lo que se usan técnicas de aprendizaje automático, como los algoritmos de agrupamiento (o *«clustering»*) en función de la similitud entre las imágenes, entre otras (Sommer and Gerlich, 2013).

1.1.3.2. Extracción de características de imágenes

Los resultados del análisis de imagen suelen estar compuestos por características morfológicas, geométricas, de intensidad y basadas en texturas (Figura 1.3).

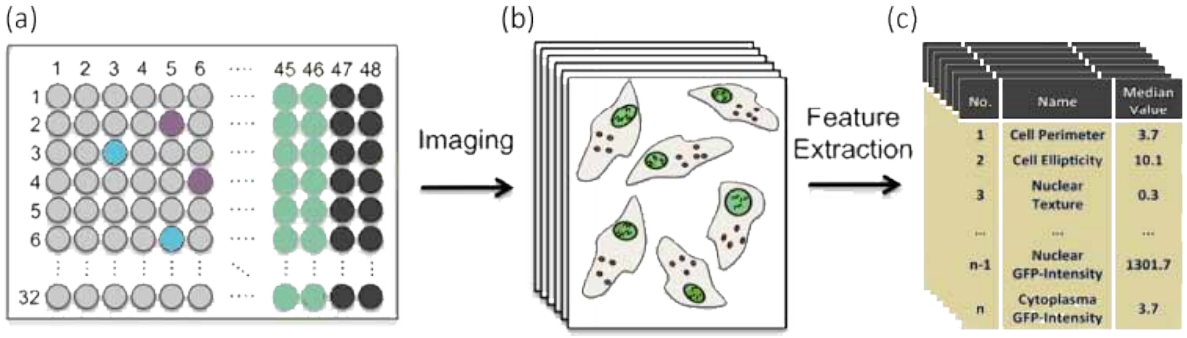


Figura 1.3: **Proceso de extracción de características. (a)** Experimento de silenciamiento génico. **(b)** Obtención de imágenes. **(c)** Extracción de características. Figura extraída de Reisen et al. (2013).

Existe una gran diversidad de herramientas para extraer características de imágenes de microscopía. Dos de ellas con amplio recorrido en el ámbito biológico son Phenoripper (Rajaram et al., 2012) y WND-CHARM (Shamir et al., 2008; Orlov et al., 2008). La primera es una aplicación de código abierto implementada en MATLAB y diseñada para facilitar tanto el análisis como la exploración de imágenes. Aunque es una herramienta muy versátil que permite comparar imágenes en base a su similitud fenotípica, el principal inconveniente que presenta es la carencia de una interfaz basada en línea de comandos, dificultando así su incorporación a flujos de trabajo para la ejecución en un entorno de supercomputación. WND-CHARM además de extraer las características de una imagen, también permite su clasificación. En este caso, sí que hay una

interfaz mediante línea de comandos y además, está parcialmente integrada en la plataforma OMERO[1].

El objetivo de la extracción de características es representar con valores numéricos el contenido de una imagen, para lo que se usa un conjunto de funciones matemáticas que completan un vector de 2659 características distintas clasificadas en:

A **Características de alto contraste** (Tabla 1.1): estadísticas sobre el número de objetos, distribución espacial, tamaño, forma, etc.

B **Descomposiciones polinomiales** (Tabla 1.2): polinomio que aproxima el contorno de formas detectadas en la imagen con cierta fidelidad y los coeficientes se usan como descriptores del contenido de las imágenes.

C **Texturas** (Tabla 1.3): variaciones de intensidad entre píxeles en distintas direcciones y resoluciones.

D **Parámetros estadísticos sobre los valores de los píxeles** (Tabla 1.4): distribución de las intensidades de los píxeles representados en histogramas, incluyendo también el cálculo de los momentos de dicha distribución (véase la Tabla 1.3).

Función	**Longitud**	**Características**
Bordes	28	Media, mediana, varianza, histograma, número total de píxeles ocupados por bordes, etc.
Objetos	34	Resultado de aplicar una máscara binaria sobre la imagen y buscar los 8 elementos conectados en la máscara binaria resultante. También se calcula el mínimo, máximo, media, mediana, varianza e histograma.
Gabor	7	Detección de texturas y patrones periódicos.

TABLA 1.1: **Grupo A. Características de alto contraste.** Representa valores estadísticos de la imagen, como el número de objetos que contiene, la distribución espacial, el tamaño, la forma, etc.

[1] www.openmicroscopy.org/site/products/partner/wnd-charm

Función	Longitud	Características
Estadísticas de Chebyshev	32 * 2	Aproximación usando un polinomio aplicado a la imagen original y a la transformada de Fourier.
Estadísticas de Chebyshev-Fourier	32 * 2	Se aplican dos transformaciones 2D en distancia y ángulo, donde la distancia se aproxima mediante un polinomio de Chebyshev y el ángulo mediante ondas de Fourier. De esta forma, se detectan grandes áreas con transiciones de intensidad suaves.
Polinomios de Zernike	72 * 2	Coeficientes de la aproximación polinomial de la imagen calculados sobre la imagen original y la transformada de Fourier.

TABLA 1.2: **Grupo B. Descomposiciones polinomiales.** La imagen es aproximada mediante un polinomio en dos variables cuyos coeficientes son representados en un vector.

Función	Longitud	Características
Cálculo de los 4 momentos	48 * 6	Media, varianza, asimetría y curtosis (grado de concentración en torno a la media) donde cada imagen se divide en franjas en 4 orientaciones distintas (0°, 90°, 45° y -45°).
Texturas de Haralick	28 * 6	Se computa sobre la imagen original y sobre 5 transformadas (Fourier, Chebyshev, Wavelet, Wavelet y Chebyshev de la transformada de Fourier).
Histograma multi-escala	24 * 6	Histogramas variando el número de cajas (3, 5, 7, 9). Se aplica a 6 imágenes distintas: la imagen original, la imagen transformada de Fourier, la imagen transformada de Chebyshev, la transformada Wavelet y la Wavelet y Chebyshev de la transformada de Fourier.
Texturas de Tamura	6 * 6	Contraste (rango de intensidades de los píxeles), tosquedad (escala de la textura) y direccionalidad (dirección) para la imagen original y las 5 transformadas.

TABLA 1.3: **Grupo C. Estadísticas y texturas.**

Función	Longitud	Características
Características del Grupo C		Incluye todas las transformaciones del Grupo C (tabla 1.3).
Transformada de Radon	12 * 4	Información espacial a través de la transformada de Radon, que calcula una proyección de las intensidades de los píxeles en una línea radial desde el centro de la imagen hasta el borde de la misma según un ángulo dado (0°, 45°, 90°y 135°).

TABLA 1.4: **Grupo D. Estadísticas, texturas e información espacial.**

Aunque se aplica un rango variado de funciones a la imagen original y en algunos casos también a las transformadas de las mismas, el espacio de valores que se genera asociado a las características no es ortogonal, lo que implica que ciertas características pueden variar de forma dependiente. Tampoco es un conjunto completo, así que puede haber aspectos de la imagen que no se analicen. Además, las características pueden presentar patrones no informativos o variables correlacionadas, por lo que puede ser necesario reducir las dimensiones. Para una descripción de los métodos de reducción de dimensiones, véase la sección 1.1.11.1.

1.1.3.3. Limitaciones de la anotación de fenotipos

El análisis automático de imágenes no es una estrategia infalible. Están descritos en la literatura algunos artefactos que pueden surgir durante la aplicación de esta técnica. Por ejemplo, en Shamir (2011), se seleccionaron experimentos ya publicados en los que se clasificaban imágenes en función de su contenido biológico. Con el objetivo de analizar la eficiencia de esta clasificación, se sustituyeron regiones de la imagen con alta intensidad (como pueden ser las células) por cuadrados de color blanco con área similar (Figura 1.4). Cuando se clasificaron de nuevo las imágenes en función de su morfología, la precisión en la predicción (47 %) fue muy similar a la obtenida con las imágenes originales (52 %). Esto se traduce en que un artefacto en una imagen podría hacerse pasar por una célula. Los histogramas con las intensidades de los píxeles tampoco son adecuados para discriminar los artefactos en este caso, puesto que la distribución de los niveles de intensidad se mantiene al sustituir la célula por el artefacto. Este experimento pone de manifiesto la importancia de que los descriptores de imágenes procedentes de microscopía sean los adecuados. De no ser así, ciertos artefactos podrían ser clasificados por el algoritmo generando falsos positivos. El análisis computacional podría, por tanto, estar sesgado por los artefactos en vez de venir guiado por su contenido biológico.

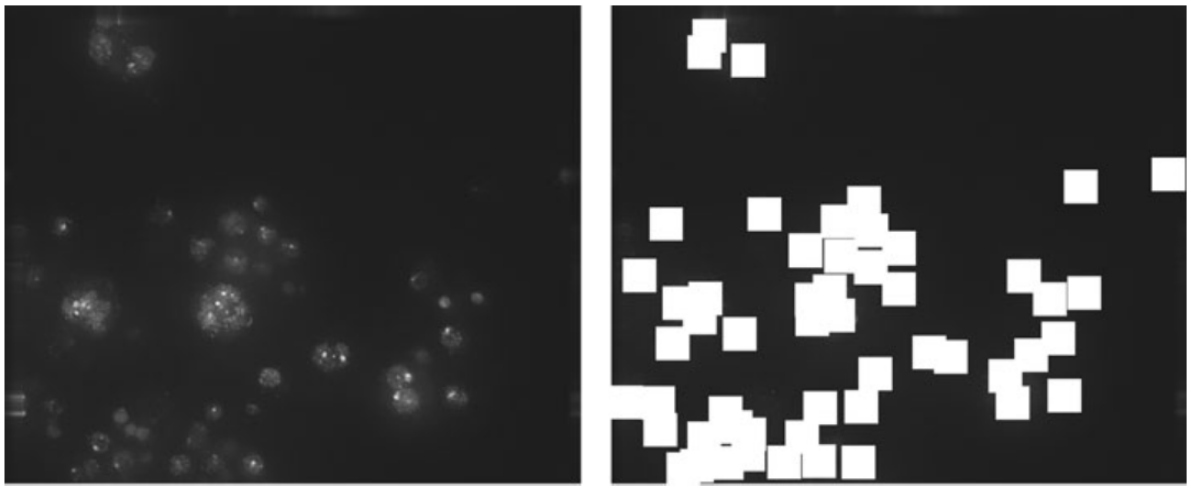


Figura 1.4: **Ejemplo de fenotipo y su correspondiente conversión en artefacto.** Ejemplo de una imagen procedente de un conjunto de datos de RNAi (izquierda) y la misma imagen modificada sustituyendo cada punto de la imagen original por un cuadrado blanco de $60x60$ píxeles (derecha). Figura extraída de Shamir (2011).

Otro sesgo importante derivado del diseño experimental es el conocido como efecto *«batch»* (Definición 9), que se da cuando las condiciones de laboratorio afectan a los resultados del experimento.

> *Definición* 9. **Efecto *«batch»*:** fenómeno por el cual un conjunto de muestras producidas conjuntamente se parecen más entre sí que aquellas producidas en lotes distintos, con independencia de las variables científicas en estudio.

Aunque este efecto no es exclusivo de los experimentos de alto rendimiento, se detecta más fácilmente cuando hay más datos y puede paliarse mediante el tratamiento computacional de los resultados (Leek et al., 2010). Entre estos métodos para paliar la heterogeneidad de lotes distintos se encuentra el *análisis de componentes principales* o PCA (*«Principal Component Analysis»*) junto con algoritmos de visualización como el *agrupamiento jerárquico* o el *«MultiDimensional Scaling»* (MDS) (véase la sección 1.1.11.1).

Ya se han descrito en secciones anteriores otras posibles inconsistencias que podrían derivarse de la intervención humana en el proceso de anotación. Esto también ocurre en el ajuste de parámetros a la hora de tomar las muestras, lo que podría hacer que una imagen se clasifique erróneamente por pequeñas diferencias en dichos ajustes en vez de por su contenido biológico.

Por todos estos motivos, es importante partir de un diseño del flujo de trabajo y de los algoritmos consistente, con la ayuda parcial que proporciona la inspección visual de las imágenes.

1.1.4. Perfiles fenotípicos

En los experimentos de alto rendimiento que generan y analizan imágenes se ha producido una transformación en la aproximación –usando la misma tecnología– para cuantificar el contenido biológico: desde la identificación de fenotipos previamente conocidos (escaneo o *«screening»*) hacia la captura de características de forma imparcial (evaluación por perfiles o *«profiling»*) (Caicedo et al., 2016).

El *«screening»* es el enfoque tradicional, donde el experimento busca cuantificar un único proceso o función celular midiendo sólo algunas características de interés. Esos fenotipos permiten elegir un subconjunto de genes para un estudio en mayor profundidad. Como se conoce de antemano el fenómeno a medir, este método depende de la experiencia previa del investigador que diseña el experimento, pues interrogará el fenómeno de una forma u otra.

El segundo enfoque se denomina *«profiling»* y en él no se estudia ningún proceso concreto, sino que se recopila un amplio rango de mediciones, sin hacer uso de ningún tipo de conocimiento previo que guíe el proceso. Desde esta perspectiva, se aborda el proceso en estudio desde una posición más distante, dotando así al resultado de un menor sesgo y de un mayor potencial para detectar mecanismos desconocidos (Caicedo et al., 2016).

En ambos enfoques, el resultado de la anotación de los fenotipos procedentes del análisis de imagen se puede concretar en un **vector de características** multiparamétrico con valores cuantitativos que describen la observación obtenida de una imagen o muestra. Por ejemplo, en el *«profiling»* morfológico se extraen grandes cantidades de mediciones morfológicas (Figura 1.5b) para describir aspectos sobre la forma, el tamaño, la intensidad, textura, etc. de distintos compartimentos celulares.

Una de las dificultades que surgen en el *«profiling»* morfológico es la complejidad de los datos resultantes. Normalmente, este tipo de experimentos suele constar de cientos de placas, cada una de ellas con 384 pocillos. De cada pocillo se obtiene al menos una imagen –pueden ser también varios fotogramas de un vídeo– y de cada imagen se obtiene un vector de características –a veces altamente correlacionadas entre sí– que refleja de forma cuantitativa los valores de distintos aspectos morfológicos de una célula (Figura 1.5). A esto hay que añadir que una muestra procedente de un pocillo contiene distintas células, por lo que se hace necesario obtener un perfil común que refleje el comportamiento general de todas las células de la muestra. La

variabilidad en un pocillo puede deberse a la adaptación dinámica de las células individuales al microentorno del experimento. Por ello, también es importante prestar atención al contexto de la célula: si se encuentra aislada, con un área de contacto grande con otras células, etc. ya que no siempre se mantiene la consistencia entre réplicas de experimentos de RNAi que inhiben el mismo gen (Snijder et al., 2012).

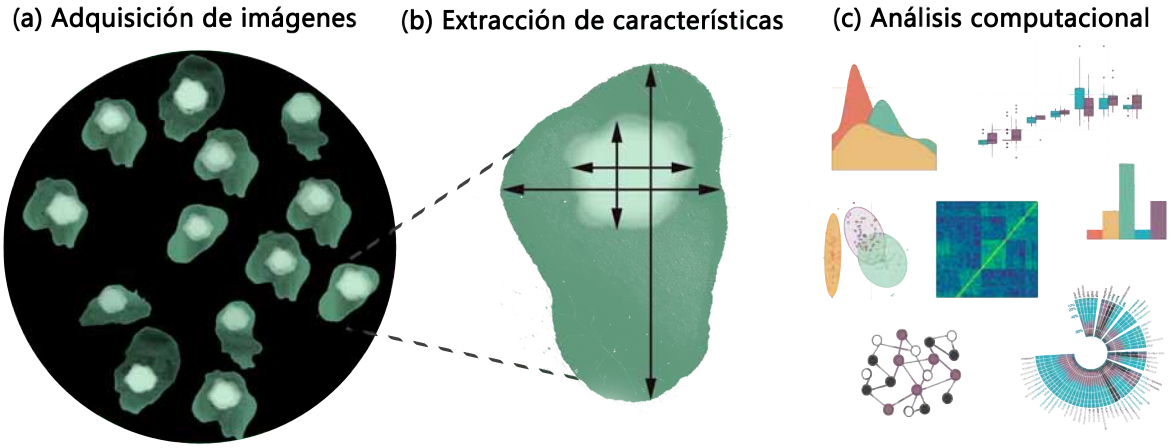


Figura 1.5: **Evaluación por perfiles para imágenes celulares. (a)** Adquisición de imágenes de microscopía. **(b)** Extracción de características. **(c)** Análisis computacional de las mediciones obtenidas.

El método más utilizado para generar un vector de características común a la muestra consiste en calcular la media de la observación para cada característica. Usar el promedio resulta muy ilustrativo cuando la población de células es homogénea, pero no siempre se dan dichas circunstancias. La variabilidad de los fenotipos es muy común en los procesos biológicos y puede deberse al ruido con respecto a un único estado fenotípico medio. Por ello, los métodos que generalizan usando la media suelen dar peor resultado que aquellos en los que se tiene en cuenta la variabilidad en las muestras u observaciones (Ljosa et al., 2013).

Una forma de abordar esta dificultad de forma computacional es considerar el espacio fenotípico de una muestra como una respuesta variada y compleja. Aplicando un método no supervisado de aprendizaje automático se pueden agrupar las células en categorías fenotípicas no definidas *a priori*. El reto en este caso aparece cuando hay que asignar un significado biológico a esos grupos, aunque en algunos trabajos ha funcionado detectando variaciones morfológicas para las etapas de la mitosis (Zhong et al., 2012).

Para sortear este inconveniente, el usuario podría definir las categorías fenotípicas de antemano, en vez de ser detectadas mediante aproximaciones no supervisadas. Una opción es tratar a las poblaciones celulares heterogéneas como si fuesen mezclas de subpoblaciones distintas fenotípicamente, donde un modelo mixto Gaussiano («*Gaussian Mixture Model*», GMM) permite agrupar las células en un espacio de características multidimensional. Así lo considera (Slack et al., 2008) en la caracterización de respuestas de células cancerosas a distintos fármacos. Otro algoritmo supervisado muy común en este campo es *Support Vector Machine* (SVM, Ben-Hur et al. (2008)), que presenta una buena capacidad de generalización, aunque para ello necesita un conjunto de entrenamiento representativo que haya sido anotado previamente de forma manual, con los inconvenientes ya mencionados que ello conlleva. Aunque esto dota de una gran facilidad de interpretación de los fenotipos, asume que cada célula debe pertenecer a una de las categorías previamente representadas, tomadas como clases discretas, lo que puede dar lugar a grandes pérdidas de información al obviar fenotipos existentes en la muestra pero no definidos en las categorías (Liberali et al., 2014).

Muchos de los problemas computacionales que aparecen podrán posiblemente resolverse en un futuro cercano gracias a las técnicas de «*deep learning*» (LeCun et al., 2015). Estos métodos se usan para resolver problemas de visión por computador y se empiezan a aplicar en el campo de las imágenes biológicas.

A pesar de todas las complicaciones ya mencionadas, son muchas las ventajas que presenta el fenotipado a través de vectores de características. Primero, se trata de un método más completo que la anotación de una única observación, ya que al ser multiparamétrico, se capturan distintas características simultáneamente. Además, cuando el vector de características consta de muchos parámetros medidos, el impacto de los efectos «*off-target*» se reduce, simplemente porque la probabilidad de que los genes afecten siempre a las características que capturan artefactos del mismo modo es baja. La forma habitual de etiquetar los experimentos suele ser puntuar las muestras con respecto a uno o varios fenotipos ya conocidos (por pertenecer a la misma ruta metabólica por ejemplo). Sin embargo, los vectores de características capturan fenotipos no conocidos previamente, lo que permite detectar respuestas celulares más sutiles. Por ello, el método es particularmente útil cuando los fenotipos no se conocen *a priori*, como ocu-

rre en la caracterización de fármacos (Perlman et al., 2004; Feng et al., 2009; Johannessen et al., 2015).

Tanto la aproximación no supervisada –agrupaciones sin significado fenotípico previo– como la supervisada –definición categorías fenotípicas de antemano– ayudan a asignar determinados fenotipos a los genes silenciados. Se puede representar esta asignación como un vector asociado a un gen cuyos valores indican cuantitativamente el nivel de aparición de un fenotipo. A este vector se le denomina **perfil fenotípico**.

1.1.5. Métodos estadísticos para el análisis de datos multidimensionales

La necesidad de la aplicación de métodos estadísticos al análisis de datos multidimensionales nace del tamaño y la complejidad de los datos a procesar en el análisis de imágenes. Si cada imagen se caracteriza con un vector de n dimensiones y un microscopio puede producir miles de imágenes de forma automática; extraer conclusiones directas no siempre es un procedimiento trivial.

Para facilitar y dar fiabilidad al estudio estadístico es fundamental seguir un diseño experimental adecuado (Malo et al., 2006). Normalmente, el objetivo de un experimento es detectar una serie de genes cuyas respuestas se espera que sean diferentes de un control negativo, o bien, que sean similares al control positivo. En el caso de los RNAi, los controles positivos son dsRNA conocidos por inhibir un cierto gen, mientras que los negativos son dsRNA diseñados de forma sintética para que no inhiban ningún gen. Lo ideal es que los controles se distribuyan aleatoriamente sobre toda la placa, pero normalmente se destina la primera columna para los controles positivos y la última para los negativos (Figura 1.6a).

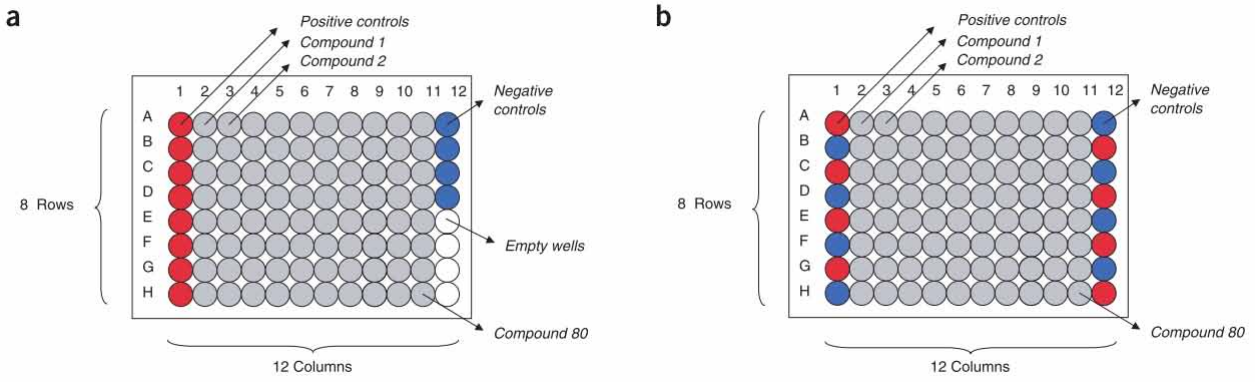


Figura 1.6: **Distribución de controles en una placa de 96 pocillos. (a)** En una placa de 96 pocillos se suelen analizar unos 80 compuestos distintos (siRNAs, fármacos, etc.), que se sitúan en la parte central de la placa (gris). En la primera columna se sitúan los 8 controles positivos (rojo) y los controles negativos (azul) ocupan las 4 primeras posiciones de la última columna (12). Las 4 posiciones siguientes se quedan vacías. **(b)** Aunque la forma ideal de distribuir los controles sería aleatoriamente sobre toda la placa, con esta distribución aleatoria sobre las columnas 1 y 12 se controlan los sesgos causados por la evaporación de líquidos en los bordes debido al gradiente de temperatura. Figura extraída de Malo et al. (2006).

Una alternativa –respetando la restricción de colocar los controles sólo en las columnas 1 y 12– consiste en colocar de forma alternada los controles positivos y negativos (Figura 1.6b), para paliar así los sesgos que pueden aparecer debidos al gradiente de temperatura de los bordes (Birmingham et al. (2009), Figura 1.7). Este efecto tiene una particular importancia en los controles de experimentos celulares, ya que el agrupamiento de células o la evaporación en ciertas áreas de la placa puede llevar a distintas condiciones de crecimiento celular (Lundholt et al., 2003). Esto puede hacer que los controles sean menos fiables, lo que repercutirá en la calidad de los resultados, ya que los valores de cada muestra son relativos a dichos controles.

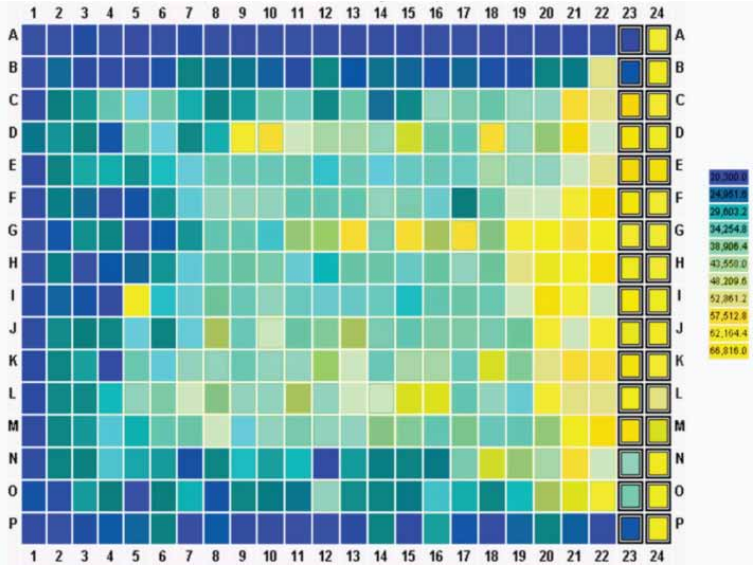


Figura 1.7: **Visualización de la salida de un experimento.** Mapa de calor correspondiente a una placa de 384 pocillos donde se muestran los valores obtenidos en un experimento hipotético sometido a los efectos del gradiente de temperatura. Figura extraída de Birmingham et al. (2009).

No obstante, aún con una distribución de controles óptima y con condiciones muy controladas, el proceso de transfección de la doble cadena exógena es una fuente de variabilidad. Además, puede causar estrés en las células y afectar a la viabilidad de las mismas, dando lugar a fenotipos indirectos como la muerte celular.

En cuanto a las similitudes entre los perfiles, los genes que se espera que manifiesten cierto fenotipo estarán cerca de los controles positivos, pues son éstos los que manifiestan los fenotipos de manera clara. La duda surge en el momento de decidir un valor umbral a partir del cual se considera un perfil significativamente similar al control positivo. Además, esta significancia

sería matemática pero realmente la que se quiere comprobar es la significancia biológica, y no siempre ambas coinciden, debido principalmente al problema de los falsos positivos (Echeverri et al., 2006).

Otro mecanismo de control de errores se centra en el uso de réplicas independientes para un mismo compuesto y bajo las mismas condiciones. Contando con un número adecuado de ellas –una práctica extendida es obtener al menos tres réplicas– se puede exigir un cierto grado de redundancia que confirme los resultados obtenidos (Echeverri et al., 2006). Aunque pueda parecer costoso o difícil de gestionar, se trata de una forma de asegurar la reproducibilidad y eliminar falsos positivos. Pero no sólo es importante la reproducibilidad de distintas réplicas, sino que también es necesario dotar al experimento de robustez en cuanto a los fenotipos obtenidos. Para ello, se suele inhibir el mismo gen con varios RNAi distintos. Una opción es hacer una segunda validación experimental con los genes identificados en la primera, así se garantiza que no ha habido ningún problema estadístico.

Al generarse tantos datos de forma automática, hay una alta probabilidad de que aparezcan artefactos ajenos al proceso en estudio. Usar esta información puede dar lugar a conclusiones incompletas o erróneas, por lo que se hace necesario un control de calidad previo al análisis estadístico, que si bien suele ser automático, depende en gran medida de la inspección visual del resultado (Hancock (2014), Chapter 5). Una herramienta útil en este caso son los mapas de calor, más comúnmente conocidos como *heat maps* (Definición 10). Representar con colores los valores de cada pocillo facilita la inspección visual para detectar de forma rápida artefactos muy notables, proporcionando una idea de lo extremo de los valores y su localización en la placa (Figura 1.7).

> *Definición* 10. **Heat map**: representación gráfica de todas las posiciones de una tabla o matriz usando colores de acuerdo a un gradiente.

Una vez eliminados todos los posibles errores que puedan introducir un sesgo en los resultados, el siguiente paso es hacer comparables entre sí los datos de distintas réplicas y/o experimentos. Para esto se usa la normalización, que homogeneiza los datos a través de distintas placas o réplicas (Figura 1.8), permitiendo también la combinación de datos procedentes de

distintas placas. Las diferencias en los datos originales pueden deberse, por ejemplo, a que las lecturas se han realizado en distintos días o por una configuración distinta del lector.

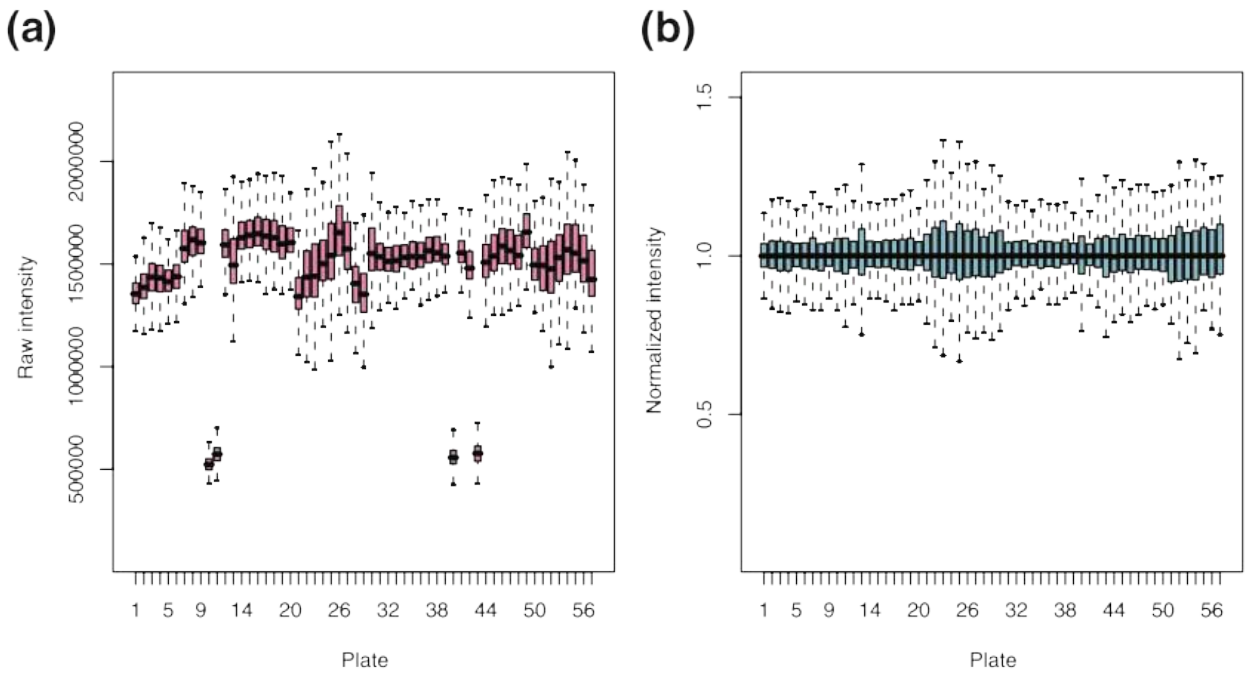


Figura 1.8: **Normalización de datos de intensidad de una placa.** Diagramas de cajas representando la intensidad de una réplica, agrupadas por placas, antes (a) y después de ser normalizadas (b). Figura extraída de Boutros et al. (2006).

Aunque hay distintos tipos de normalización (Malo et al. (2006), Box 1), una de las más comunes cuando hay controles negativos y positivos es el *Z-score*, que se define como el número de desviaciones estándar desde la media:

$$Z = \frac{x_i - \overline{x}}{s_x}$$

donde x_i es la medida de la i-ésima placa, $\overline{x}$ es la media de todas las placas y s_x la desviación típica de la medida en todas las placas.

Una ventaja del *Z-score* es que incorpora información de las variaciones en la medida de la muestra, pero también depende del uso de los controles negativos en su cálculo. Como el *Z-score* es sensible a los valores atípicos de la distribución, es más recomendable usar, para datos de RNAi, una variante conocida como el *Z-score* robusto:

$$Z = \frac{x_i - med}{MAD}$$

donde med es la mediana de la distribución a través de las placas y MAD la desviación absoluta de la mediana (MAD, *Median Absolute Deviation*).

1.1.6. Medidas de similitud vectorial

Tras el silenciamiento de un gen, el análisis de la imagen que se obtiene genera un vector de características. Estos vectores se combinan para integrar las distintas réplicas y recoger la variabilidad a nivel poblacional (más detalles en la sección 1.1.4). Aplicando aprendizaje automático estos vectores de características se traducen en perfiles fenotípicos asociados a genes. En este trabajo, los perfiles fenotípicos se considerarán vectores binarios donde se registra, para un gen dado, la presencia o ausencia de un conjunto de fenotipos.

En definitiva, con la creación del perfil fenotípico se obtiene información sobre la función de un gen. La asunción tras el uso de los perfiles es que fenotipos similares deben tener causas también similares. En este caso, al analizar experimentos sobre la pérdida de función usando RNAi, se asume que hay una relación funcional entre los genes que presentan el mismo fenotipo (Neumann et al., 2010). Se asume también que los perfiles fenotípicos de proteínas que interactúan deben ser más similares fenotípicamente que aquellos cuyas proteínas no lo hacen (Fuchs et al., 2010).

Para medir la similitud entre perfiles fenotípicos existe una gran variedad de métricas. En este trabajo se ha implementado y comparado un conjunto de medidas de distinta naturaleza para tal fin (Tabla 1.5). La selección de una métrica adecuada es un prerrequisito para el análisis de los datos derivados de HCS (Definición 7) y en todo caso, la selección de una u otra dependerá de las características del conjunto de datos biológicos a analizar (Reisen et al., 2013). Es importante destacar que, dado que los perfiles son multidimensionales, las distancias o similitudes han de medirse en un espacio de las mismas dimensiones, que coincidirán con el número de fenotipos en estudio.

Tipo	**Métricas**
Grupo A: Basada en normas	Euclídea, Minkowski, City Block, Manhattan, Mahalanobis
Grupo B: Basada en ángulos	correlación, coseno
Grupo C: Cadenas binarias	Hamming, Jaccard
Grupo D: Ponderación de términos	Cohen's Kappa, TF-IDF

TABLA 1.5: **Medidas de similitud vectorial agrupadas por categorías.** La fórmula de cada métrica se encuentra en la tabla 4.4, dentro del capítulo de Material y Métodos.

Todas las métricas basadas en normas (Tabla 1.5, Grupo A) son equivalentes según el teorema *Theorem 6.1.5* de (Conway, 2012). Entre estas métricas están: *Euclídea*, *Minkowski* (idéntica a la anterior pero para exponente 2), *City Block* o *Manhattan* (mide la distancia entre dos puntos siguiendo una cuadrícula) y *Mahalanobis* (sigue una distribución χ^2). Como todas generan valores proporcionales, aquí usaremos únicamente la similitud *Euclídea* como representante de la clase.

Aunque la distancia *Euclídea* es una de las más usadas, un problema común es el conocido como fenómeno de concentración de distancias (Kabán, 2012). Esta limitación se da cuando las dimensiones del perfil son grandes, por lo que agrupar perfiles puede resultar problemático: al aumentar el número de dimensiones, la distancia entre los puntos tiende a una constante, es decir, la distancia entre dos perfiles muy cercanos es 0 pero también la distancia entre dos perfiles muy lejanos es 0. Este es un serio inconveniente en espacios multidimensionales para ciertas métricas. Pero este no es este el único problema de los espacios con alta dimensionalidad, sino que puede darse el caso de que los perfiles no sean comparables ni interpretables (Zimek et al., 2012). La solución al problema de concentración suele ser, o bien la aplicación de una medida distinta –como el *coseno*, que se concentra pero a una tasa más baja (Radovanović et al., 2010), esto es, para perfiles con un mayor número de dimensiones– o bien aplicar un método de reducción de dimensiones para eliminar redundancias antes de comparar los perfiles (véase la sección 1.1.11.1).

Otras métricas (Tabla 1.5, Grupo B) se basan en ángulos, como es el caso de la aplicación del *coseno* o de la *correlación de Pearson* a los vectores fenotípicos.

Si en lugar de considerar los perfiles como vectores espaciales se tratan como cadenas binarias (Tabla 1.5, Grupo C), esto eso, de ceros (ausencia del fenotipo) y unos (presencia del fenotipo), entonces es posible aplicar métricas como *Hamming* o *Jaccard*. La más simple de ellas es *Hamming*, una similitud comúnmente usada en computación que mide el número de fenotipos coincidentes sobre el total de coincidencias entre dos perfiles. En la similitud de *Jaccard* se varía ligeramente la aproximación: se calcula la proporción de elementos similares sobre los disímiles, sin tener en cuenta los valores nulos.

Una representación visual de la obtención de los perfiles y los valores que presentan algunos pares de genes para las métricas hasta ahora descritas se encuentra en la Figura 1.9.

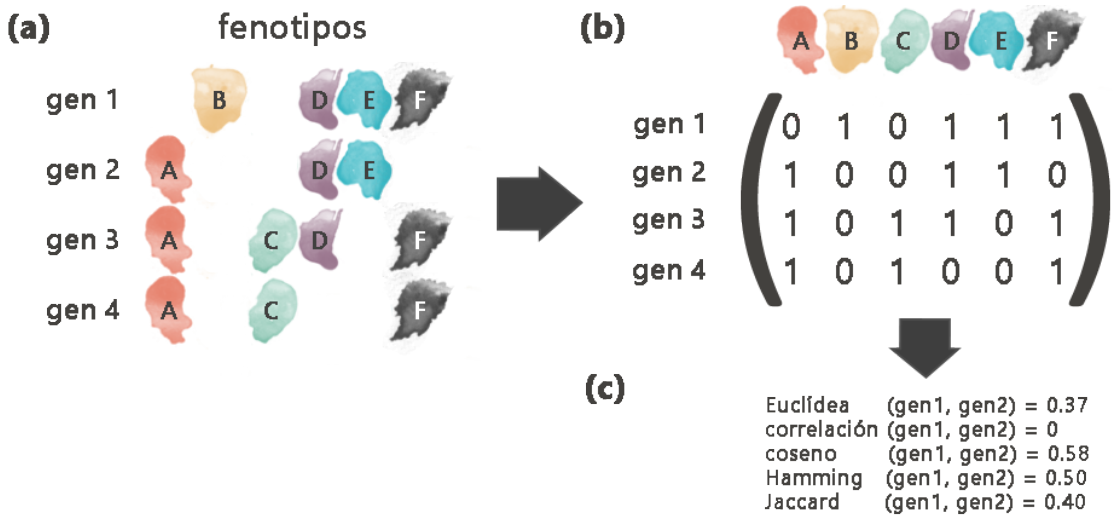


Figura 1.9: **Perfiles fenotípicos para métricas de similitud vectorial. (a)** Tras el silenciamiento de 4 genes se manifiesta en total un conjunto de 6 fenotipos (A-F). Por ejemplo, tras la inhibición del gen 1, se observan los fenotipos B, D, E y F. **(b)** Para cada uno de los genes se obtiene un perfil binario que representa la presencia (1) o ausencia (0) de un fenotipo dado. El resultado de todos los perfiles es una matriz binaria. A partir de esta matriz ya pueden calcularse las distancias entre todos los posibles pares de genes. **(c)** A modo de ejemplo, aquí se muestran los valores para el par (gen1, gen2).

Uno de los motivos por el que algunas de estas métricas no suelen aportar precisión sobre el tipo de datos que se manejan se debe a que miden las coincidencias entre los perfiles. Esto significa que cuando dos genes inhibidos manifiestan el mismo fenotipo, se podría decir que hay un acierto y que los perfiles son un poco más cercanos. Sin embargo, ocurre lo mismo en el caso de que ambos perfiles no presenten el fenotipo en común. Por ejemplo, en la Figura 1.9, el *gen 1* y el *gen 2* no manifiestan el fenotipo *C*, lo que hace los perfiles algo más cercanos al haber una coincidencia en la ausencia del fenotipo. Esto es una desventaja, porque al tener en cuenta el número de ausencias en común, se está considerando que una disimilitud es igual de relevante que una similitud. Además, esto es impreciso, pues muchas veces los fenotipos simplemente no se anotan. Cabe también la posibilidad de que el perfil provenga de experimentos distintos con fenotipos seleccionados con distintos enfoques. Por ejemplo, hay experimentos que sólo se basan en la observación de la topología celular, ignorando cualquier otra disfunción que pueda aparecer. No significa, por tanto, que dichos fenotipos no hayan aparecido al inhibir los genes,

sino que no se han tenido en cuenta por no ser de interés para el estudio. Por este sesgo en el enfoque a la hora de anotar fenotipos, surge la necesidad de aplicar métricas más elaboradas para extraer la similitud de la forma más fiel posible.

Una opción es considerar los perfiles fenotípicos como cadenas de información para medir sus similitudes y diferencias (Tabla 1.5, Grupo D). En este sentido, se usan aquí distintas métricas basadas en la teoría de la información (Shannon, 1948): *TF-IDF («Term Frequency - Inverse Document Frequency»)* (Robertson, 2004) y *Cohen's kappa*. En estas dos similitudes se ponderan los fenotipos en función de lo informativos que son: aquellos fenotipos más comunes tendrán un contenido informativo menor –se consideran fenotipos más generales– mientras que aquellos que aparecen en baja frecuencia asociados al silenciamiento de un pequeño número de genes –fenotipos específicos– tienen un contenido informativo mayor.

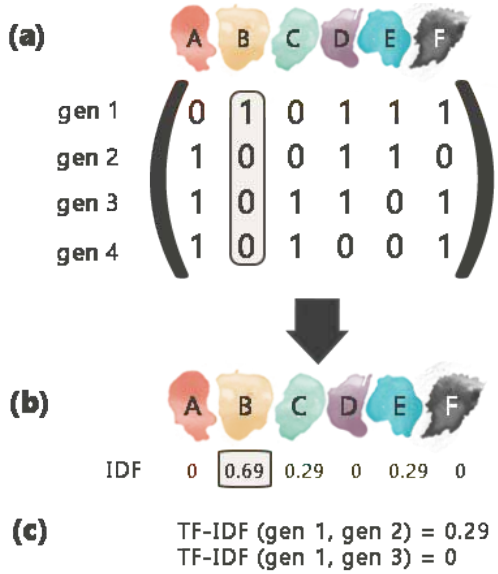


Figura 1.10: **Métricas de similitud TF-IDF. (a)** Partiendo de la matriz binaria que representa los perfiles fenotípicos de los genes 1-4, se calcula el valor de *IDF* **(b)** para cada fenotipo. De esta forma queda indicado el fenotipo más específico. **(c)** La similitud entre perfiles fenotípicos de genes se calcula usando el máximo valor de *IDF* cuando ambos perfiles son coincidentes en 1 (presencia del fenotipo). Esto se traduce en que se selecciona el valor del fenotipo más específico si ambos genes muestran dicho fenotipo. Por ejemplo, los genes 1 y 2 manifiestan conjuntamente los fenotipos D y E. Los *IDFs* asociados a esos fenotipos son, respectivamente, 0 y 0.29. Por tanto, el mayor valor de *IDF* para ambos perfiles es 0.29. Análogamente, para el par de genes 1 y 3, el fenotipo más específico que comparten es D, que presenta un *IDF* de 0.

De acuerdo con la teoría de la información, la métrica de similitud *TF-IDF* puntúa cada fenotipo con un valor que indicará su especificidad, que se traduce en un valor de IDF («*Inverse Document Frequency*»). Este valor será muy bajo si el fenotipo es muy frecuente o muy alto si aparece sólo tras el silenciamiento de muy pocos genes. Luego, teniendo en cuenta cuándo dos perfiles manifiestan el mismo fenotipo, se selecciona el mayor *IDF* de entre los fenotipos coincidentes (véase la Figura 1.10).

Otra opción para el cálculo de similitudes ponderando los términos es el coeficiente de *Cohen's kappa*. El valor de similitud entre dos perfiles se calcula como la tasa de coincidencia entre dos perfiles, o lo que es lo mismo, el número de fenotipos sobre el total que dos genes manifiestan (sin tener en cuenta el sesgo por azar). Esta es una pequeña variación sobre la versión original de *Cohen's kappa*, que incluye en el cómputo a aquellos fenotipos que ninguno de los dos genes manifiestan. El motivo de esta variación viene dado por el hecho de que un gen no esté anotado en un fenotipo no significa necesariamente «no fenotipo», sino que no fue observado, o no fue analizado por estar fuera del foco de interés del estudio, entre otros.

1.1.7. Ontologías

En la sección 1.1.6 se han detallado algunas métricas de similitud vectorial entre perfiles fenotípicos. En los vectores descritos, los fenotipos son considerados conceptos independientes, es decir, no están organizados de forma estructurada. Sin embargo, esta falta de relación entre fenotipos no ha de darse necesariamente, ya que muchos de ellos presentan relación biológica: participan en un mismo proceso, la función que desempeñan está relacionada, aparecen en el mismo orgánulo o incluso describen el mismo fenómeno.

La estandarización de los términos fenotípicos permite el procesamiento automático de los datos y la comparación entre distintas aproximaciones. Sin embargo, la estandarización no refleja las relaciones entre dichos fenotipos. Por ejemplo, dados los fenotipos *«chromosome segregation defect»* y *«metaphase arrest»*, se podría crear una categoría que describa un fenómeno mitótico que englobe a ambos. La nueva categoría estaría vinculada a los fenotipos mencionados mediante relaciones –pueden ser de distinta naturaleza– que permiten formalizar una estructura sobre la que aplicar algoritmos de razonamiento que generen nuevo conocimiento. Esta forma de organizar la información –junto con otras características– es lo que se conoce como una ontología (Definición 3).

De forma técnica, una ontología incluye distintos **recursos**: descripciones de *clases* (Definición 11), *instancias* (Definición 12) y *propiedades* (Definición 13) para modelar un **dominio** o área de interés.

> *Definición* 11. **Clase o término**: tipo de objeto que comparte definición en un dominio. Es la unidad básica dentro de una ontología.

> *Definición* 12. **Instancia**: individuo perteneciente a una única clase.

> *Definición* 13. **Propiedad o atributo**: característica o parámetro con el que se caracteriza una clase. Pueden ser de tipo dato –*«data properties»*– (código, nombre alternativo, descripción, etc.) o bien de tipo objeto –*«object properties»*– que referencie a otras instancias. A través de este último tipo de atributo, las clases se organizan jerárquicamente, siendo una **superclase** su antecesora y una **subclase** su descendiente.

Un ejemplo visual de los distintos recursos de una ontología –junto con la forma de razonamiento automático que la estructura ontológica permite– se encuentra en la Figura 1.11. Con el razonamiento no sólo se infieren relaciones sino que se puede verificar la consistencia de la ontología. El modelo que se usa aquí es conocido como *Entity-Quality* (EQ) (Mungall et al., 2010), donde se indica la entidad y la característica que la describe.

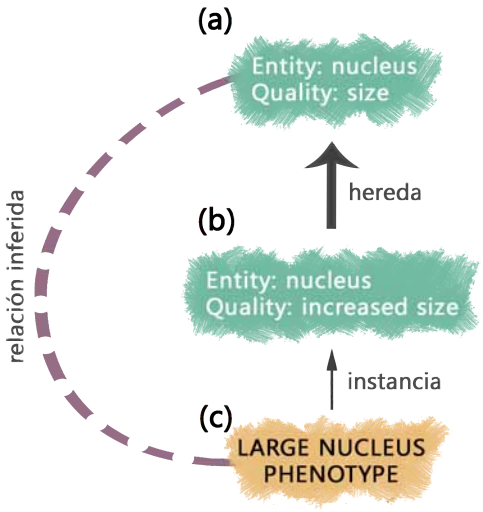


Figura 1.11: **Ejemplo de razonamiento sobre una ontología.** Se representan en verde dos clases, **(a)** y **(b)**, de una ontología ejemplo y una instancia **(c)** de la clase **(b)**, en amarillo. La clase **(a)** define características relacionadas con el tamaño del núcleo celular, mientras que la clase **(b)** representa el tamaño incrementado del núcleo. La clase **(a)** es antecesora de **(b)** y **(b)** hereda de **(a)**. Dado un nuevo fenotipo observado **(c)**, que instancia la clase **(b)**, al razonar sobre la estructura ontológica, se infiere automáticamente que **(c)** también es instancia de **(a)**. De esta forma se ponen de manifiesto relaciones que no son inmediatamene obvias en la ontología. Ejemplo extraído de Dessimoz and Škunca (2017).

Las ontologías biológicas usadas en este trabajo difieren ligeramente de la definición que se aplica en el campo de la informática. En primer lugar, la nomenclatura cambia: las clases ahora son **términos** (Definición 11) y desaparecen las instancias y los atributos. En su lugar, aparecen las **anotaciones** (Definición 14). Estas anotaciones son, por lo general, genes o productos de genes que manifiestan el comportamiento descrito por uno o varios términos (como se indicó en la Definición 12, en el caso de las instancias esta multiplicidad no es posible).

> *Definición* 14. **Anotación**: asociación entre los términos y los genes o productos de genes (proteínas, RNA no codificante, complejos macromoleculares u otros).

Aunque técnicamente las ontologías biomédicas no suelan cumplir los requisitos para ser ontologías, en el ámbito biológico se sigue usando este término para describir un **vocabulario controlado** (Hoehndorf et al., 2015). El objetivo es organizar la información para dotar de un estándar que defina los términos y las relaciones entre ellos, de forma que se represente el dominio en estudio. Al solucionar las variaciones en la terminología se facilita la comunicación, interoperabilidad, integración, acceso a los datos y análisis de los mismos (Soldatova and King, 2005).

En el ámbito biológico y biomédico, las ontologías se expresan en un lenguaje formal, normalmente *OWL* (*«Ontology Web Language»*), que se basa en la lógica de predicados de primer orden, siendo *OBO* (*«Open Biomedical Ontology»*) el formato tradicional en que se han representado las ontologías, aunque ambos formatos son interconvertibles (Tirmizi et al., 2011). Otro lenguaje basado en grafos es *RDF* (*«Resource Description Framework»*), que representa la información en forma de tripletes siguiendo el formato `<sujeto, predicado, objeto>`.

Entre los repositorios de ontologías más importantes se encuentran *Ontology Lookup Service*[2] (Côté et al., 2006), *BioPortal*[3] (Noy et al., 2009), *OBO Foundry*[4] (Smith et al., 2010) y *OntoBee*[5] (Xiang et al., 2011).

La aplicación más directa de las ontologías consiste en dotar de un esquema sobre el que construir una base de datos. Pero sin duda, la funcionalidad más útil en el contexto biológico es la organización jerárquica proporcionada por las relaciones entre términos.

[2] http://www.ebi.ac.uk/ols
[3] http://bioportal.bioontology.org
[4] http://www.obofoundry.org
[5] http://www.ontobee.org

1.1.7.1. Organización funcional de genes

Una de las ontologías más conocidas en la clasificación funcional de los genes es *Gene Ontology* (GO) (Ashburner et al., 2000), que almacena y recoge información sobre los genes en tres niveles o **aspectos** distintos (Figura 1.12): el proceso biológico en que participan (Definición 15), la localización en la que actúan (Definición 16) y la función molecular que desempeñan (Definición 17). Además, estandariza la representación de genes procedentes de distintas especies y bases de datos, lo que facilita la interoperabilidad y la comparación. Aunque de aquí en adelante se use el término ontología, ya se ha explicado en la sección 1.1.7 que esto es una imprecisión, pues a nivel técnico es un vocabulario controlado (Smith et al., 2003). La relación entre los genes y los términos o clases a las que pertenecen se encuentra en el fichero de anotaciones (GOA, *Gene Ontology Annotation*) (Camon, 2004) separadas según el aspecto concreto de la ontología:

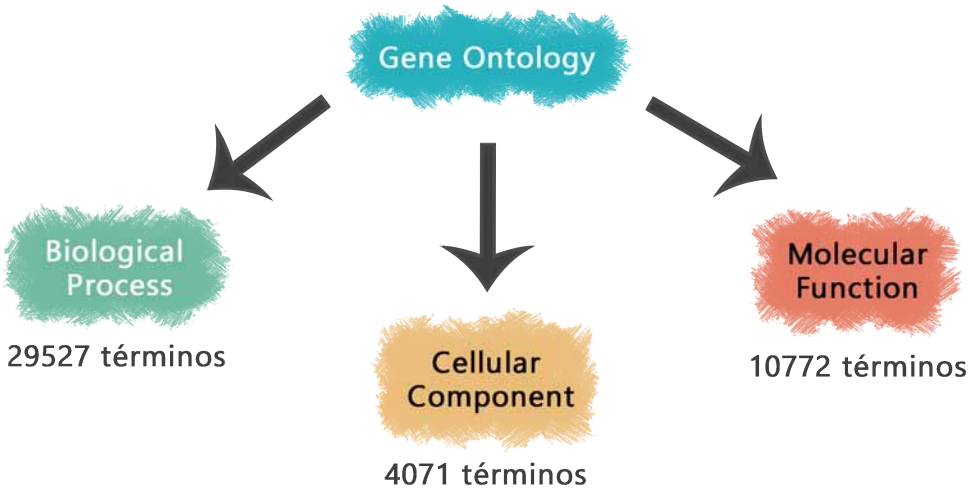


Figura 1.12: **Ramas de la ontología *Gene Ontology*** junto con el número de términos para cada subontología de GO, a fecha 6 de Febrero de 2017.

> *Definición* 15. **Procesos biológicos *(BP, Biological Process)***: evento o agrupación de eventos moleculares con inicio y fin definidos en los que un gen participa, relativos al funcionamiento de las células, tejidos, órganos y organismos. Estos procesos pueden involucrar una transformación química o física.

> *Definición* 16. **Componentes celulares *(CC, «Cellular Component»)***: partes de la célula o del espacio extracelular donde un gen está activo.

> *Definición* 17. **Función molecular *(MF, «Molecular Function»)***: actividad bioquímica de un gen, sin indicar el momento y lugar en que ocurre.

Se puede considerar cada uno de estos tres dominios como ontologías independientes –no son redundantes ni comparten raíz común– aunque sí hay relaciones entre ellos, principalmente entre los procesos biológicos y las funciones moleculares. Las tres ramas comparten un espacio común de identificadores y una sintaxis propia. La estructura es un grafo acíclico dirigido (DAG, *«Direct Acyclic Graph»*, Definición 18) donde los nodos representan los términos.

> *Definición* 18. **Grafo acíclico dirigido *(DAG, «Direct Acyclic Graph»)***: conjunto de nodos conectados a través de aristas dirigidas que representan las relaciones entre ellos con ausencia total de ciclos dirigidos. Un nodo puede tener uno o más nodos padres y cero o más nodos hijos.

Las relaciones entre términos pueden tener distinta naturaleza (Figura 1.13): *«is a (is a subtype of)»*; *«part of»*; *«has part»*; *«regulates»*, *«negatively regulates»* y *«positively regulates»*. La estructura básica de GO se fija sobre relaciones *«is a»* (Figura 1.13a) para definir subtipos de términos, siendo ésta además la relación más frecuente en todos los aspectos de la ontología GO. Las relaciones de tipo *«part of»* (Figura 1.13b) se usan para definir partes de un todo, como por ejemplo, etapas de un proceso biológico, siendo la relación simétrica *«has part»* (Figura 1.13c). Finalmente, las relaciones de regulación (*«regulates»*) (Figura 1.13d, e y f) describen la influencia de un proceso sobre otro. Estas relaciones son propias de GO y reflejan la influencia específica de un proceso sobre otro.

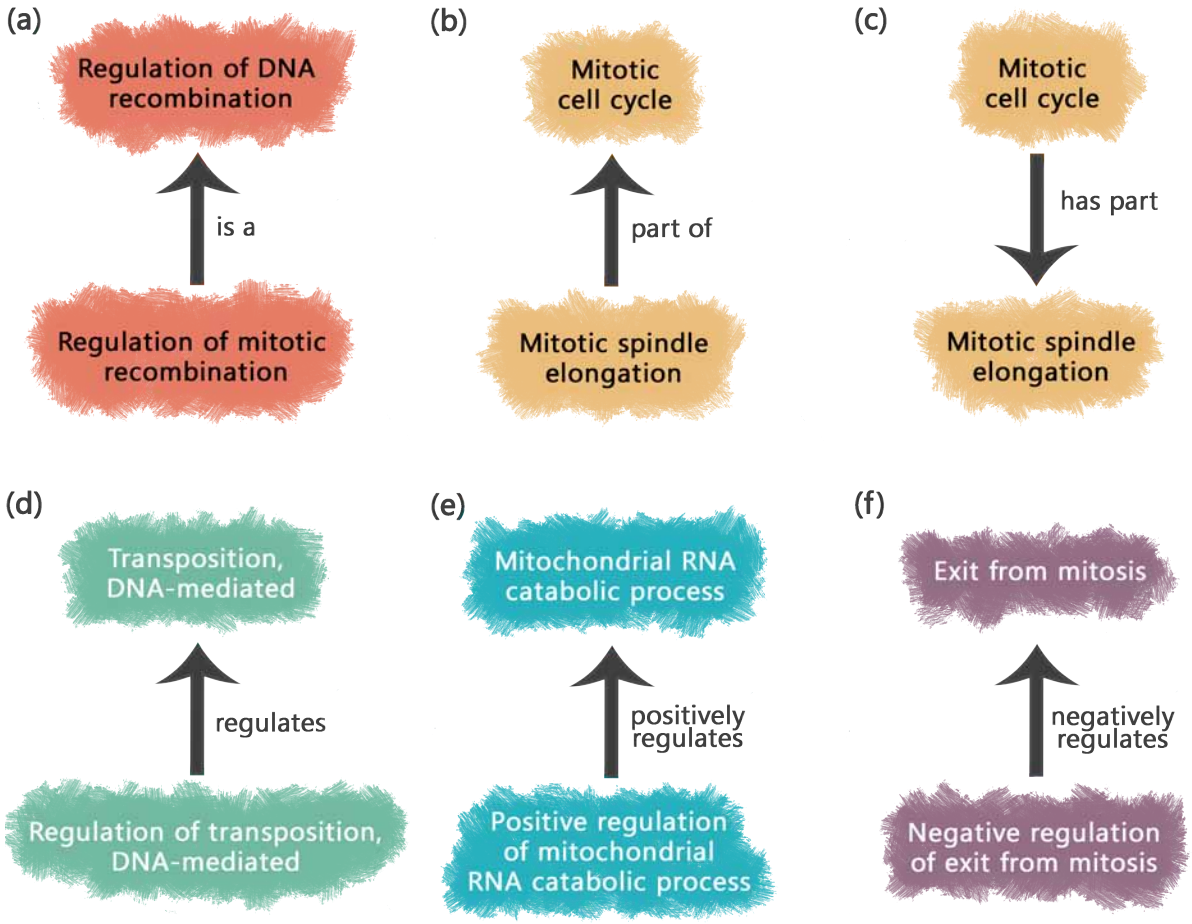


Figura 1.13: **Tipos de relaciones en GO.** La relación *«is a»* conforma la estructura básica de GO. Las relaciones *«part of»* y *«has part»* se refieren a la misma relación, son simétricas y dan nombre distinto dependiendo de la dirección. Por último, las relaciones de regulación reflejan la influencia de un término sobre otro.

Una vez definida la estructura del grafo formada por los términos y relaciones de distinta naturaleza, es necesario definir los tipos de relaciones entre genes y los términos que los clasifican. Dependiendo de la evidencia que respalde la anotación, se define un conjunto de códigos[6] que aquí se explicarán agrupados por categorías:

- ✧ **Experimental**: la relación se extrae de un artículo científico que caracteriza físicamente un gen. Hay varios códigos experimentales, aunque el que se refiere a funciones obtenidas a partir de la mutación de un único gen es `IMP` (*«Inferred from Mutant Phenotype»*).

[6] `http://geneontology.org/page/guide-go-evidence-codes`

- ✧ **No experimental o computacional**: la anotación se sustenta en un análisis *in silico*, normalmente derivado del estudio de la secuencia del gen. Cuando se asocia un término a un gen por métodos automáticos sin supervisión humana se usa el código `IEA` (*«Inferred from Electronic Annotation»*).

- ✧ **Declaración de autor en una publicación**: la anotación se hace en base a una afirmación del autor, bien sea porque se cita en un artículo de investigación, revisión o entrada de una base de datos.

- ✧ **Curado**: cuando el análisis del artículo no deriva en ninguna de las categorías anteriores se anotan como `IC` (*«Inferred by Curator»*) o bien `ND` (*«No biological Data available»*). Se basan en el conocimiento de un experto a partir de un contexto de datos pero sin que haya evidencia directa disponible.

Para cuantificar el número de evidencias de cada tipo que hay en las anotaciones de genes de *«Homo Sapiens»* en términos GO, se han extraído los códigos del fichero de anotaciones GOA y se ha representado su frecuencia en la Figura 1.14. Un gen puede estar anotado al mismo término con más de una evidencia.

Figura 1.14: **Distribución de evidencias en GO para *Homo Sapiens*.** El grupo de evidencias más común es el experimental, seguido por el computacional y los declarados por el autor. Los datos curados conforman la minoría de anotaciones.

El uso de GO está muy extendido para anotar genes obtenidos a través de las técnicas *«ómicas»* (Definición 2). La clasificación de los genes –atendiendo a su comportamiento molecular, localización o proceso en que participan– hace que los datos sean fáciles de interpretar de forma sistemática. Por ejemplo, se pueden buscar genes que compartan características (localización, proceso, etc.); estudiar si en un grupo de genes procedente de un experimento aparecen algunos términos sobrerrepresentados; inferir la función de genes que no se han anotado aún; o, como suele hacerse en los experimentos de alto rendimiento, determinar los procesos en que difieren dos conjuntos de genes.

La topología en forma de grafo facilita la medición de distancias entre nodos, para lo que son necesarias nuevas métricas que se detallarán en la sección 1.1.8.

1.1.7.2. Organización de fenotipos celulares

Gracias a la gran aceptación de las ontologías en la comunidad biomédica, éstas se han establecido como el método estándar para la representación de fenotipos (Oellrich et al., 2016). También en el ámbito de la biología celular se han desarrollado en los últimos años algunas iniciativas para estandarizar la información fenotípica que se obtiene en los ensayos de alto rendimiento.

Un primer intento fue la ontología **CPO (*«Cell Phenotype Ontology»*)** (Hoehndorf et al., 2012), generada automáticamente a partir de GO. Esta ontología clasifica anomalías morfológicas y fisiológicas de células, componentes y procesos celulares. La estructura sobre la que se basa es la de *«Phenotypic Attribute and Trait Ontology»* (PATO), que surgió en el año 2002 con el objetivo de capturar la información fenotípica en organismos muy diversos. La idea subyacente es que todas las descripciones tengan una entidad (enzima, proceso biológico, estructura anatómica, etc.) y un atributo (cuantitativo o cualitativo) que la describa. Los atributos suelen ser la duración, localización temporal de un proceso, el comienzo, la frecuencia, etc. Por ejemplo, son fenotipos: *«alcohol dehydrogenase is absent»*, *«cell division is arrested at metaphase»*, *«eyes are absent»* o *«hyperactive»*. Este mecanismo de composición de fenotipos es el modelo *«Entity-Quality»* (EQ), que se representó en el ejemplo de la Figura 1.11. Si el fenotipo está descrito como un concepto atómico (un único término que engloba la definición de la entidad y su característica) habría que desgranarlo en el formato EQ:

Concepto:
«Erythrocytopenia»
Entity:
«Red blood cells»
Quality:
«Deficiency»

En el caso de la ontología CPO, el identificador de cada fenotipo en la ontología representa tanto la entidad como la característica siguiendo el patrón `CPO:QQEEEEEEE`, donde `QQ` representa la cualidad y el resto de dígitos (`EEEEEEE`) la entidad sobre la que se aplica, que en esta ontología es un término GO. Por ejemplo:

Clase GO:

«Apoptosis» (`GO:0006915`)

Clases CPO derivadas:

«Abnormality of apoptosis» (`CPO:120006915`)

«Abnormality of single occurrence of apoptosis» (`CPO:140006915`)

«Abnormality of regulation of apoptosis» (`CPO:150006915`)

Uno de los inconvenientes que presenta esta ontología es su tamaño: describe cada uno de los términos de GO aplicando distintos calificativos, lo que hace que el número de términos multiplique el de GO, haciendo poco viable tanto la búsqueda de términos como la anotación de genes, además de dificultar su mantenimiento y futura ampliación.

Como evolución natural de esta ontología, y cambiando de forma radical el enfoque, surge recientemente **CMPO (*«Cellular Microscopy Phenotype Ontology»*)** (Jupp et al., 2016), que describe y organiza fenotipos celulares a partir de datos experimentales, con anotaciones manuales. La principal diferencia con CPO es que viene guiada por resultados, lo que la hace más compacta, sencilla y directa.

En esta ontología, además de describir algunos fenotipos para las tres ramas de GO (definiciones 19, 20 y 21), también se incluyen descripciones sobre células individuales y poblaciones (definiciones 22 y 23). A diferencia de GO, aquí sí hay una raíz común, por lo que se considera una única ontología (Figura 1.15). Las relaciones entre términos son de la misma naturaleza que las de GO.

> *Definición* 19. **Proceso celular** ***(«Cell Process Phenotype»)***: descripciones fenotípicas en el ámbito de los procesos celulares.

> *Definición* 20. **Componente celular** ***(«Cellular Component Phenotype»)***: descripción de características que afectan a componentes celulares de GO.

> *Definición* 21. **Componente molecular** ***(«Molecular Component Phenotype»)***: fenotipos a nivel molecular de la célula usando como entidades las biomoléculas descritas en la ontología ChEBI (Degtyarenko et al., 2008).

Definición 22. **Célula individual** ***(«Single Cell Phenotype»)***: fenotipos observados a nivel de célula completa.

Definición 23. **Población** ***(«Cell Population Phenotype»)***: conjunto de fenotipos propios de una población de células.

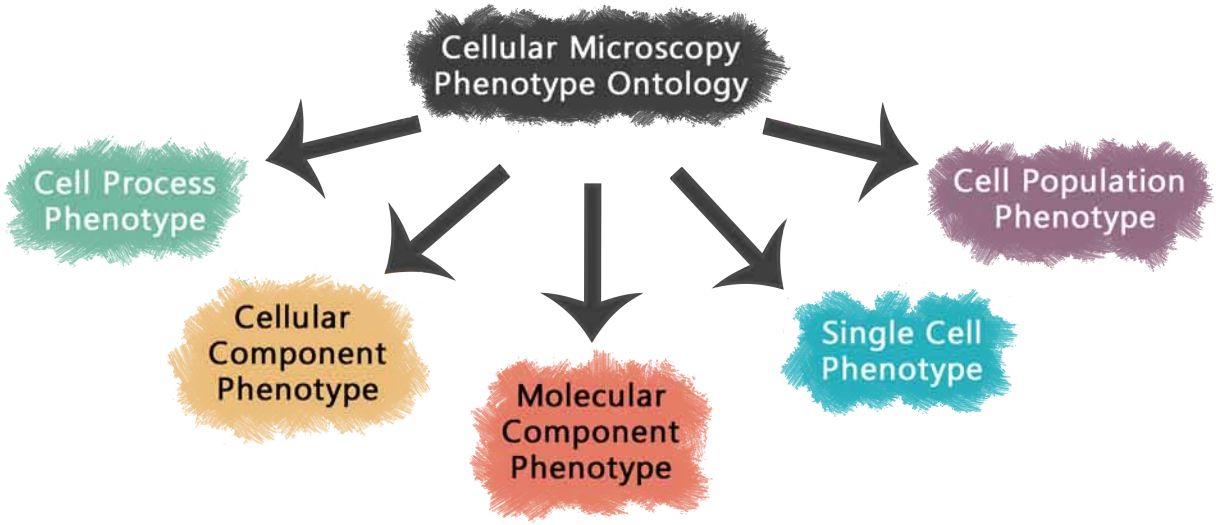


Figura 1.15: **Principales ramas de la ontología CMPO.** Por un lado se describen fenotipos relacionados con las categorías principales de GO: *biological process*, *cellular component* y *molecular function*. Además, hay una categoría para la descripción de fenotipos en células individuales («*Single Cell Phenotype*») y otra para la descripción de poblaciones («*Cell Population Phenotype*»).

CMPO se usa actualmente para anotar fenotipos que aparecen en los datos del «*Image Data Repository*» (IDR, Williams et al. (2017)), incluido en la plataforma OMERO, que se mencionó en la sección 1.1.3.2.

1.1.8. Medidas de similitud en ontologías

En la sección 1.1.6 se describieron distintas métricas para calcular distancias entre perfiles fenotípicos. Estas métricas tomaban los fenotipos como distintas posiciones de un vector de dimensiones igual al número total de fenotipos. Gracias a la organización de fenotipos en una estructura ontológica, aparecen nuevas alternativas para medir la similitud, por estar esos fenotipos relacionados entre sí.

La forma en que se organizan los términos en una ontología proporciona numerosas ventajas. Una de ellas es que al tratarse de un grafo, se pueden calcular distancias entre pares de términos basadas en la topología (Rada et al., 1989). La distancia conceptual entre dos términos puede definirse como el mínimo número de aristas que los separan. Por ejemplo, en la Figura 1.16, los términos *«pyramidal neuron migration»* e *«interneuron migration»* están al mismo nivel y separados por dos aristas en el grafo. De forma análoga, la distancia entre *«locomotion»* y *«growth»* también es dos, pero conceptualmente son términos mucho más lejanos que el par anterior. Para un mismo valor de distancia, se observan, por tanto, pares con distinto grado de similitud a nivel semántico.

Hay otras variantes que se centran en calcular el promedio de todos los caminos entre dos términos o seleccionar el más corto. No obstante, todos estos métodos basados en la topología no tienen en cuenta la posición del término en el grafo: puede verse en la Figura 1.16 que los nodos en la parte superior de la jerarquía representan conceptos más generales que los que ocupan la parte inferior.

Conforme se desciende en la ontología, las descripciones de los términos van siendo más específicas a nivel biológico. Pero también la especificidad depende de los genes anotados en un término. En la mayoría de experimentos de RNAi diseñados para estudiar un proceso biológico en concreto, la muerte celular es posiblemente el fenotipo más fácilmente identificable y candidato a ser anotado, aunque no aporta mucha información más allá de que el gen es esencial para el funcionamiento de la célula. Este fenotipo sería frecuente y poco informativo sobre el experimento, ya que puede deberse a multitud de factores, como se indicó en la sección 1.1.5.

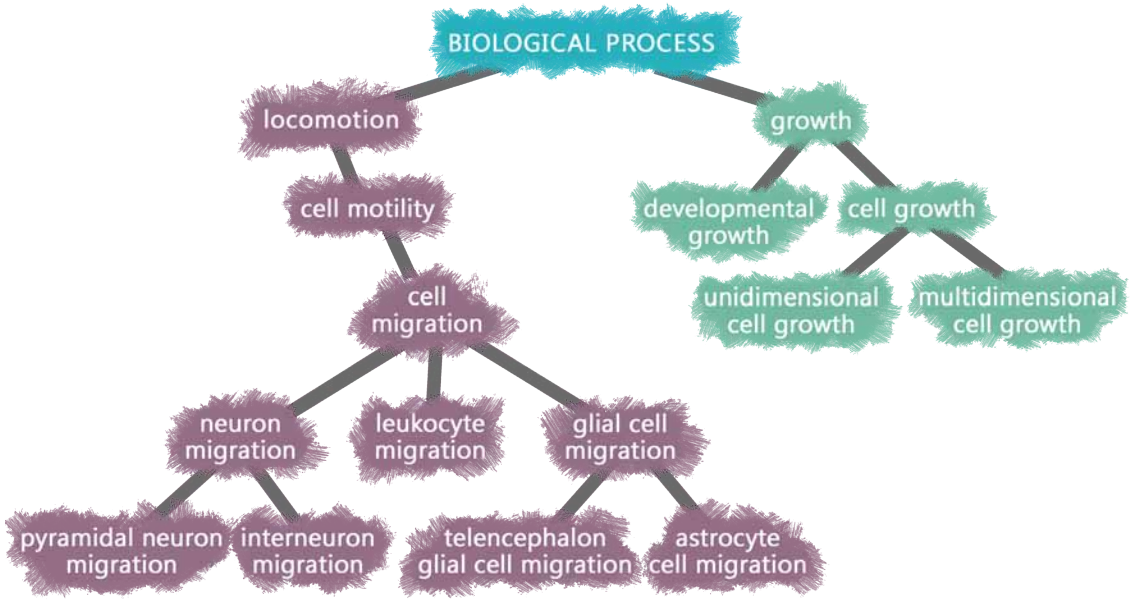


Figura 1.16: **Representación de dos ramas de la ontología GO a modo de ejemplo.** Se han seleccionado algunos términos procedentes de dos ramas de GO-BP – *«locomotion»* y *«growth»*–, con sus correspondientes relaciones jerárquicas. Aunque, por simplicidad, en la figura las relaciones entre términos no aparecen como aristas dirigidas, la relación real es de tipo *«is a»* donde cada término apunta a su ascendiente en la jerarquía. Midiendo la distancia desde una aproximación topológica, los términos *«pyramidal neuron migration»* e *«interneuron migration»* están separados por dos aristas en el grafo. Por otro lado, la distancia entre dos términos muy generales como son *«locomotion»* y *«growth»* también es dos, aunque conceptualmente sean términos muy distintos.

Siguiendo con la Figura 1.16, el término *«leukocyte migration»* tendrá menos genes anotados que *«cell migration»*, ya que este último reunirá todas las anotaciones de sus descendientes. Esto se produce por transitividad y por la inferencia de las anotaciones explicada en la Figura 1.11. Aplicando este principio a todos los términos, la raíz –*«Biological Process»*– contendrá todos los genes anotados. Por tanto, la probabilidad de que un gen esté anotado en el término raíz es uno. Y esa probabilidad irá decrementándose al descender en la jerarquía (Figura 1.17). Es importante destacar aquí que para el cálculo de las anotaciones sólo se tienen en cuenta las relaciones de tipo *«is a»* y se obvia el resto (Figura 1.13).

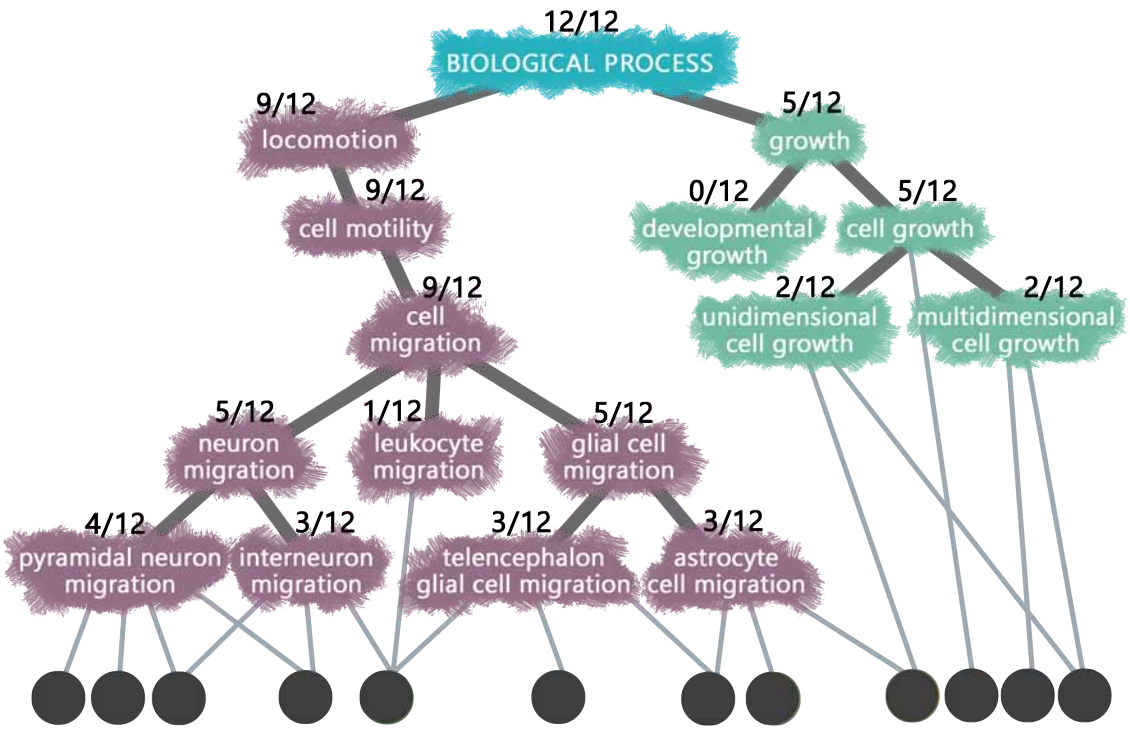


Figura 1.17: **Ejemplo de anotaciones en una ontología.** Los genes anotados se representan con círculos grises en la parte inferior y sobre cada término aparece la probabilidad de que, en general, un gen esté anotado en dicho término. Por ejemplo, en el término *«pyramidal neuron migration»* hay 4 genes de los 12 anotados en toda la ontología. Al inferirse las relaciones de los genes con términos superiores en la ontología se tiene que el término *«Biological Process»* tiene anotados la totalidad de los genes (12/12).

El **contenido informativo (IC, «Information Content»)** o **valor semántico** de un término –relacionado con su grado de especificidad semántica– se define como:

$$IC(t) = -log(p(t))$$

siendo $p(t)$ la fracción de genes anotados en el término t (o sus descendientes) en la ontología, o lo que es lo mismo, la probabilidad de que un gen esté anotado en el término t, basándose en la teoría de la información de Shannon (Shannon, 1948).

Este parámetro establece la especificidad de un único término, pero para usarlo en la medición de similitudes es necesario definir un criterio que establezca dicha similitud entre un par de términos. Generalmente, la forma más sencilla consiste en calcular el IC de un ancestro común. Una opción consiste en seleccionar el ancestro común a dos términos que se sitúe más cercano a ambos en la ontología, lo que se conoce como el ancestro común más específico

o **MICA («Most Informative Common Ancestor»)** (Resnik, 1995). La idea intuitiva es que dos términos que comparten un ancestro común muy específico (alto *IC*) deben ser muy similares.

En la Figura 1.17, el *MICA* de los términos *«pyramidal neuron migration»* e *«interneuron migration»* es *«neuron migration»*, por ser el término más específico de todos los antecesores comunes hasta la raíz. El *IC* del *MICA* es $-log(5/12) = 0{,}38$. Por el contrario, para el par de términos *«cell motility»* y *«cell growth»*, el término común más específico que los une –su *MICA*– es *«Biological Process»*, con un $IC = -log(12/12) = 0$. En estos dos ejemplos se observa que cuando se tiene en cuenta la distribución de las anotaciones y la estructura de la jerarquía, se puede estimar con mayor precisión la similitud semántica entre términos.

Por todo lo anteriormente expuesto, se define la **similitud semántica (SS, «Semantic Similarity»)** entre dos términos como la medida de la proximidad entre ellos dentro de la ontología teniendo en cuenta la teoría de la información sobre las anotaciones.

Hay una gran variedad de métricas para calcular la similitud semántica entre términos ontológicos (Tabla 4.5) (Pesquita et al., 2009). La más simple y conocida es *Resnik* (Resnik, 1995), que se basa justamente en el cálculo del *IC* del *MICA* como medida de similitud semántica, aunque presenta una particularidad: no se tiene en cuenta la distancia entre el *MICA* y los términos cuya distancia se mide.

Otras similitudes como *Lin* (Lin, 1998) y *Jiang* (Jiang and Conrath, 1997) sí consideran la distancia entre el término *MICA* y los dos términos entre los que se mide la similitud semántica, aunque este método puede ser controvertido en algunos casos. Por ejemplo, en la Figura 1.17, si se mide la distancia de *Jiang* entre el par términos *«astrocyte cell migration»* y *«glial cell migration»*, el *MICA* sería *«glial cell migration»*, por lo que la distancia entre ambos términos y el *MICA* sería 0. Este valor se mantiene independientemente de la posición que ocupe el par de términos en la ontología: para los términos *«locomotion»* y *«biological process»*, el *MICA* sería *«biological process»*, con una distancia entre los términos y el *MICA* de 0 también. El fenómeno que se da en estos dos ejemplos puede ser problemático en casos como éste de términos muy genéricos, pues presentarían una similitud muy alta. Como solución al problema mencionado y tomando como base la similitud de *Lin*, la similitud de *Schlicker* (Schlicker et al., 2006) incluye un factor para considerar la probabilidad de anotación en el *MICA*, incorporando así la posición en la jerarquía. Por tanto, ni *Schlicker*, ni *Pesquita* ni *Resnik* presentan este problema.

Nombre	Fórmula
Similitud semántica de Resnik (Resnik, 1995)	$s(t_1, t_2) = IC(t_{MICA})$ donde: - el ancestro común más informativo (MICA) es $t_{MICA} = argmax_{t \in S(t_1,t_2)} IC(t)$, - el contenido informativo (IC) de un término t es $IC(t) = -log(p(t))$, - la probabilidad de un término t es $p(t) = \frac{anotaciones(t)}{anotacionesTotales}$, y - $A(t_1, t_2)$ es el conjunto de ancestros comunes de t_1 y t_2.
Similitud semántica de Jiang (Jiang and Conrath, 1997)	$s(t_1, t_2) = 1 + 2 \cdot IC(t_{MICA}) - (IC(t_1) + IC(t_2))$
Similitud semántica de Lin (Lin, 1998)	$s(t_1, t_2) = \frac{2 \cdot IC(t_{MICA})}{IC(t_1) + IC(t_2)}$
Similitud semántica de Schlicker (Schlicker et al., 2006)	$s(t_1, t_2) = \frac{2 \cdot IC(t_{MICA})}{IC(t_1) + IC(t_2)} \cdot (1 - p(t_{MICA}))$
Similitud semántica de Pesquita (Pesquita et al., 2007)	$s(t_1, t_2) = \frac{\sum_{t \in A(t_1,t_2)} IC(t)}{\sum_{t \in P(t_1,t_2)} IC(t)}$ donde: - $P(t_1, t_2)$ es el conjunto de ancestros comunes de t_1 ó t_2.

TABLA 1.6: **Medidas de similitud semántica en ontologías.** Todas las métricas de la tabla se basan en el contenido informativo de los términos que se deriva de las anotaciones.

Por último, la similitud semántica propuesta por *Pesquita* (Pesquita et al., 2007) se basa en el índice de Jaccard, es decir, considera los *ICs* de los términos en común sobre el *IC* del total de ancestros de ambos términos (más detalles sobre las métricas en la Tabla 4.5).

Hasta este punto, el cálculo de la similitud semántica se ha aplicado a los términos de la ontología. No obstante, la entidad de interés biológico son los genes anotados en dichos términos, lo que permite extraer conclusiones sobre la relación entre dos genes dependiendo de la similitud entre el (los) término(s) en que se anotan. Aunque la similitud semántica entre genes se puede medir también por grupos (Dessimoz and Walker (2016), Chapter 12), aquí se usará la similitud semántica entre pares de genes como medida de su cercanía en la ontología, ya sea a escala funcional o fenotípica. Como normalmente cada gen está anotado en más de un término, la similitud entre genes se traduce en medir la similitud entre dos conjuntos de términos (Pesquita et al., 2007), o lo que es lo mismo, medir la similitud entre todos los posibles pares de términos en que están anotados dichos genes (Figura 1.18).

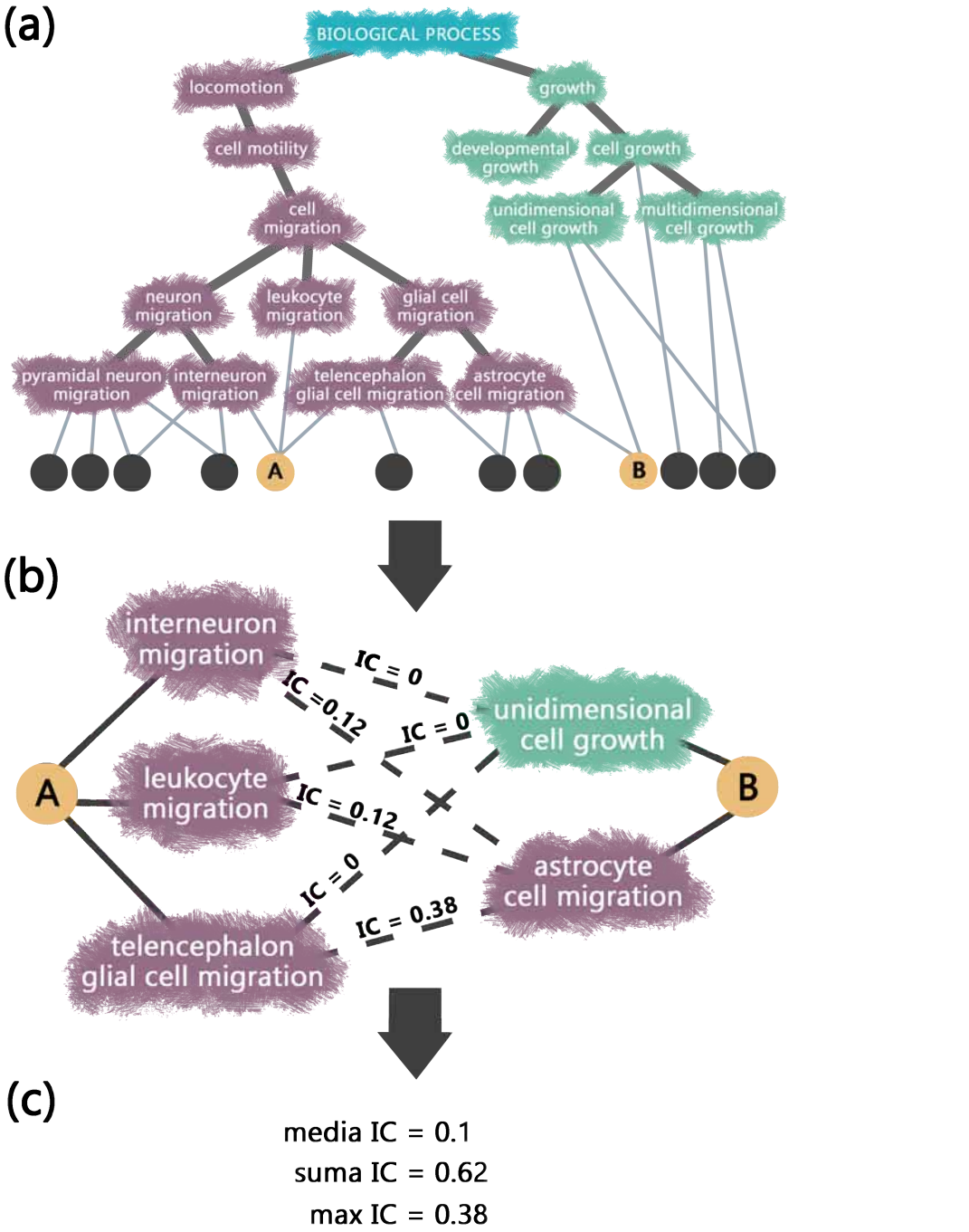


Figura 1.18: **Similitud semántica de Resnik entre los genes A y B. (a)** El gen A está anotado de forma directa en 3 términos y el gen B en 2 términos. **(b)** La similitud semántica entre los genes A y B se traduce en la similitud semántica entre todos los pares de términos (línea discontinua), siguiendo alguna de las métricas que aparecen en la Tabla 4.5. **(c)** Una vez calculadas las similitudes entre todos los pares, se combinan los valores usando la media, el máximo o la suma de las similitudes semánticas entre pares de términos.

Una vez calculada la similitud semántica entre todos los posibles pares de términos, es necesario combinar los resultados, para lo que se puede usar la media, la suma de los valores o la selección del máximo (Figura 1.18c). La combinación usando la media o la suma de todos los valores penaliza a aquellos genes que tienen múltiples funciones, ya que se va a enmascarar el efecto de aquellos términos muy similares por los que no lo son (Dessimoz and Walker (2016), Chapter 12). Es por ello que la selección del máximo suele ser la estrategia más adecuada.

A modo de ejemplo, al calcular la similitud entre genes que tienen conjuntos de anotaciones muy heterogéneos en cuanto a especificidad, se obtendrán distintos pares de términos GO donde algunos coincidirán y su similitud será alta, pero la mayoría de las combinaciones de términos serán entre pares que en principio no presentan alta similitud, debido a que el gen desempeña funciones muy distintas o manifiesta fenotipos muy diversos. Esto podría ocurrir entre los genes A y B de la Figura 1.19, que tienen anotaciones idénticas pero a términos de ramas muy distintas. Si se mide la similitud semántica de Resnik entre A y B ($s(A,B)$), se mediría la distancia entre dos pares de términos: «*unidimensional cell growth*»-«*unidimensional cell growth*» (similitud 0.78); «*astrocyte cell migration*»-«*unidimensional cell growth*» (similitud 0); y «*astrocyte cell migration*»-«*astrocyte cell migration*» (similitud 0.6). Al calcular la media, se obtendría: $s(A,B) = 0{,}35$.

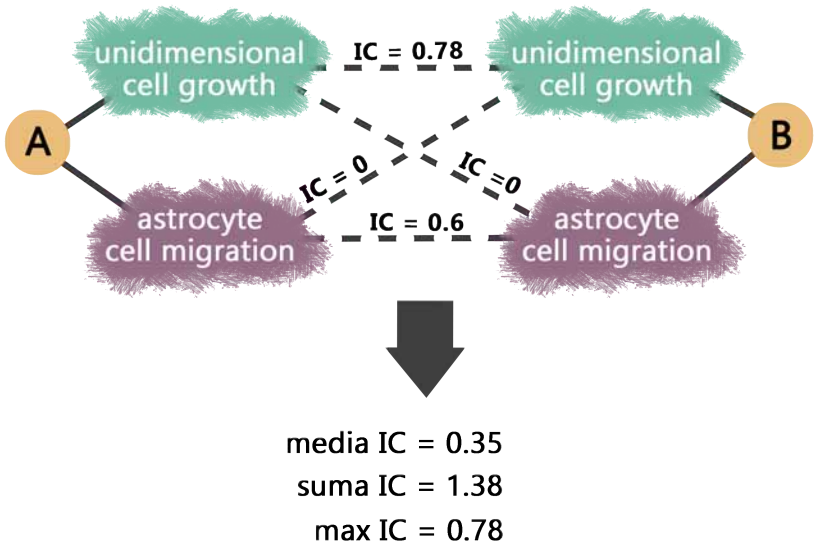


Figura 1.19: **Similitud semántica de Resnik entre dos genes con la misma anotación.** Los genes A y B están anotados en los mismos dos términos. La combinación de las similitudes semánticas entre cada par de términos siguiendo la estrategia de la suma y de la media penalizan el resultado final, mientras que el máximo parece ser la estrategia más adecuada en estos casos.

Cuando se usa el máximo para combinar los resultados de similitud entre términos, este fenómeno desaparece: siguiendo con el ejemplo anterior, ahora la similitud sería 0.78 ($s(A,B) = 0{,}78$). Sí podría darse la situación en que dos genes con anotaciones muy dispares coincidan solamente en un término cuyo IC sea máximo, y por tanto se obtenga un valor de similitud muy alto. Esto puede verse como una ventaja si se tiene en cuenta que es probable que una proteína interactúe con otra solamente por tener en común una función muy específica. También cuando hay anotaciones erróneas la aproximación del máximo podría dar lugar a errores.

Por todo lo expuesto anteriormente, la selección de una métrica adecuada a los datos es una etapa fundamental. En general, Resnik para los términos combinado con la estrategia del máximo para los genes suele ser una de las mejores (Guo et al., 2006; Xu et al., 2008; Jain and Bader, 2010; Mazandu and Mulder, 2014). Aún así, su elección no deja de ser subjetiva y a veces arbitraria, por lo que una buena aproximación puede ser aplicar varias medidas para demostrar que los datos no son sensibles a la selección de la misma (du Plessis et al., 2011).

1.1.9. Limitaciones del uso de ontologías

Tal y como se han definido aquí las ontologías biológicas, se pueden considerar herramientas muy útiles para especificar categorías y expresar axiomas, dadas las limitaciones que el lenguaje natural tiene para describir de forma precisa y estandarizada cierta información (Bodenreider and Stevens, 2006). Sin embargo, no deja de ser una forma de modelar la realidad partiendo de nuestra percepción subjetiva de la biología (Glass and Girvan, 2015). Además, al estar basadas en reglas lógicas, la representación de cierta información puede complicarse: es el caso de los datos no categóricos –como la frecuencia de aparición de un fenotipo–, de la definición difusa de términos –efecto de un fármaco en el tratamiento de una enfermedad– y de la definición de estados cambiantes en el tiempo mediante clases, como el caso en que dos genes se asocian a un término sólo en un corto espacio temporal (du Plessis et al., 2011; Dessimoz and Walker, 2016).

Aunque las medidas de similitud semántica son más informativas por tener en cuenta tanto la topología del grafo como la distribución de las anotaciones, esto también puede acarrear una serie de inconvenientes. Todas las dificultades que aquí se presentan han sido detectadas en GO por ser la ontología más usada, pero puede hacerse extensible a otras (Dessimoz and Walker (2016), Chapter 12).

Las limitaciones aparecen principalmente en dos aspectos distintos: por un lado, la forma en que se anotan genes en la ontología y por otro, cómo esto afecta a las métricas de similitud semántica.

Debido al interés variable que suscitan distintos campos de estudio biológicos, las anotaciones no suelen estar uniformemente distribuidas. Esto se conoce como *«corpus bias»* (Mistry and Pavlidis, 2008). El efecto es la distribución variable del contenido informativo en ciertas ramas, dando lugar a términos con un valor alto de IC que a la vez son poco específicos por no estar en ramas comúnmente estudiadas.

También es destacable el procedimiento por el que se anotan los genes en la ontología, ya que es necesario aportar una evidencia que dé respaldo a dicha anotación (detalles en la sección 1.1.7.1). Esta evidencia es importante por ser relevante para el cálculo de similitudes; concretamente, las anotaciones inferidas electrónicamente (IEA), al no estar curadas, suelen considerarse menos fiables. La forma más sólida en que se puede apoyar la anotación es la experimental.

Dependiendo de la estrategia usada para la combinación de términos –selección del máximo, media o suma–, y de la proporción de anotaciones IEA en los datos sobre los que se aplica, se puede obtener un resultado más o menos preciso. Si un gen está anotado con más de una evidencia a un término, a efectos del cálculo de la similitud semántica contabiliza como un único vínculo entre el gen y el término, así que a veces estas evidencias inferidas electrónicamente podrían ser solamente redundantes. En la mayoría de casos, incluirlas suele tener un efecto positivo o nulo sobre los resultados (Guzzi et al., 2012).

A la hora de anotar un gen, es posible que los datos de los que se disponga no vayan más allá de una descripción muy general. Cuando esto sucede, lo habitual es anotarlo en términos muy generales de la ontología, provocando así que dos genes que desempeñan funciones muy distintas terminen anotados en un mismo término genérico. Cuando se calcule la similitud entre dichos genes, el solapamiento entre los términos en que se anotan será completo. Este hecho también repercute en el resultado de la similitud semántica (Mistry and Pavlidis, 2008).

Dado que una ontología es una estandarización en crecimiento continuo, es habitual la definición de nuevos términos –a veces poco justificado–, lo que provoca un crecimiento lineal de ramas sin expansión horizontal, donde términos padres e hijos comparten anotaciones y por tanto tienen el mismo IC. Una forma de controlar este tipo de crecimiento es declarando obsoletos algunos términos y reemplazándolos por su antecesor (Mazandu and Mulder, 2012).

1.1.10. Redes

El cálculo de similitudes –tanto fenotípicas como funcionales–, que se ha explicado en las secciones 1.1.6 y 1.1.8, está orientado a la obtención de un valor asociado a un par de genes, ya sea mediante el uso de una ontología o a través de sus perfiles fenotípicos. Esta similitud se puede expresar en forma de matriz cuadrada, donde cada posición (X, Y) contiene el valor de similitud entre el gen X y el gen Y.

Otra forma de expresar estas similitudes es mediante una **red** o **grafo** (Definición 24) donde los nodos son los genes y el valor de similitud representa el peso de las aristas.

> *Definición* 24. **Grafo:** representación de las relaciones de un conjunto de nodos o vértices (V), donde la conexión entre nodos se indica con una arista (A). Matemáticamente, se expresa con el par $G = \{V, A\}$.

Un grafo es dirigido si las aristas tienen dirección, como sucede en el caso de la ontología GO, que es además acíclico (Definición 18). La ontología se compone de un conjunto de términos conectados a través de las aristas que indican una relación semántica entre ellos.

Cuando un grafo consta de dos tipos de nodos donde las aristas solamente unen nodos de tipos distintos, se le denomina **grafo bipartito**. Aunque esta disposición de los nodos contiene información más precisa, lo más común es comprimir la red de forma que solamente permanezcan nodos de un solo tipo. Para ello, se realiza la **proyección** de la red (Zhou et al., 2007) en dos posibles direcciones, dependiendo del tipo de nodo que conforme la red final. La proyección supone transformar la red para que contenga un único tipo de nodo (Figura 1.20). La forma habitual de obtener la proyección consiste en conectar los nodos de la red resultante cuando contienen al menos un nodo vecino –perteneciente al otro grupo– en común.

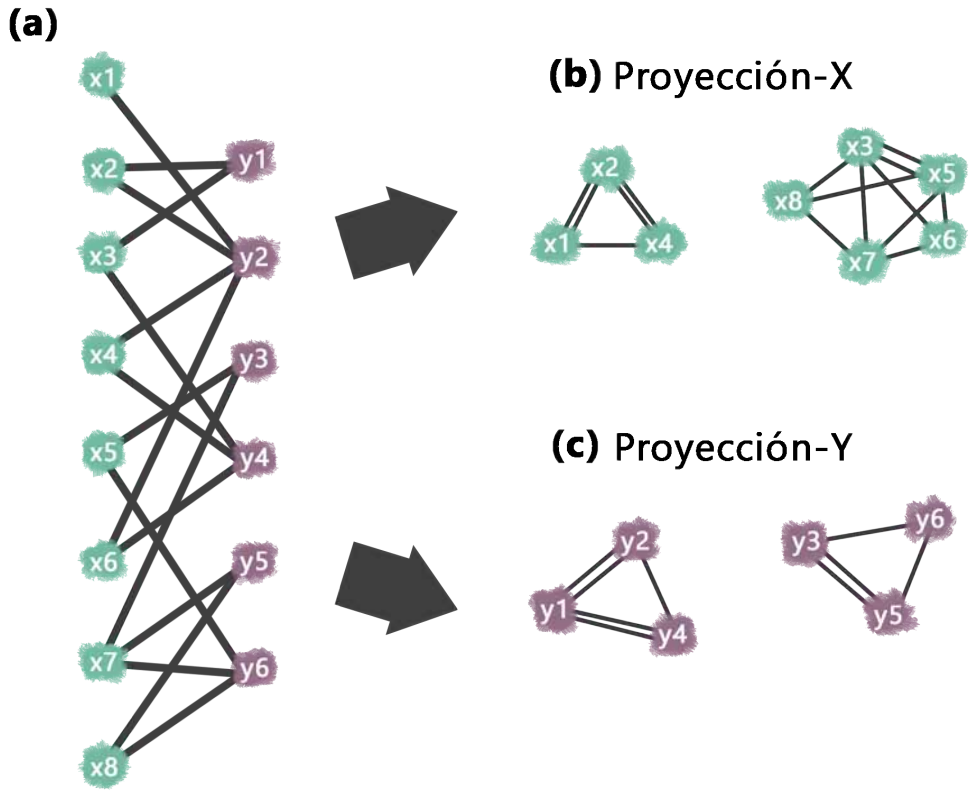


Figura 1.20: **Ejemplo de proyección de una red bipartita. (a)** La red bipartita está formada por nodos de tipo X e Y. Las dos posibles proyecciones constan de nodos de tipo X en **(b)** y de tipo Y en **(c)**. El peso de las aristas viene dado por el número de vecinos comunes en Y y X, respectivamente. Figura extraída de (Zhou et al., 2007).

Un ejemplo de grafo lo forman las proteínas que interaccionan físicamente, donde éstas son los nodos y las aristas presentan valores binarios indicando si el par conectado interactúa o no.

1.1.10.1. Redes de interacción de proteínas

Las interacciones entre proteínas son críticas en todos los procesos celulares. La función derivada de la interacción depende del correcto ensamblaje de las proteínas para convertirse en complejos funcionales, como es el caso de las maquinarias de traducción, transcripción, señalización celular (Westermarck et al., 2013), etc.

En este contexto surge el *«interactoma»* como el conjunto de interacciones físicas entre moléculas de una célula, también llamadas PPI (*«Protein-Protein Interactions»*, Definición 25).

> *Definición* 25. **Interacción proteína-proteína (PPI, *«Protein-Protein Interaction»*)**: contacto físico entre dos proteínas como resultado de un evento bioquímico que puede dar lugar a un nuevo complejo o a una interacción puntual para actividades de transporte o modificación entre ellas (fosforilación, glicosilación, metilación, etc.).

Entre los métodos más populares para la obtención de estas interacciones se encuentran Y2H (*«Yeast two-Hybrid»*) y AP/MS (*«Affinity Purification/Mass Spectrometry»*).

La técnica Y2H –descrita por primera vez en 1989 (Fields and Song, 1989)– permite la detección de proteínas interactuantes en células vivas de levadura (Brückner et al., 2009). El mecanismo se basa en la activación de un factor de transcripción que inicialmente se encuentra separado en dos partes, cada una unida a una de las dos proteínas que se estudian. Si éstas interactúan, ambos fragmentos del factor de transcripción se fusionan y se inicia la transcripción de un gen marcador que notifica que la interacción ha tenido lugar.

En el caso de AP/MS (Dunham et al., 2012), se caracterizan complejos multiproteína. Para ello se purifica el complejo, previamente identificado con una etiqueta de afinidad. Posteriormente, se digiere el complejo con una proteasa y se identifican sus componentes mediante espectrometría de masas.

Ambas técnicas captan interacciones complementarias por ser de distinta naturaleza (Yu et al., 2008): Y2H detecta las interacciones binarias directas entre pares de proteínas mientras que AP/MS identifica miembros de complejos estables. En general, la mayoría de las interacciones publicadas se han detectado usando Y2H, por ser AP/MS una técnica con mayor coste. Uno de los inconvenientes de estas técnicas es que las proteínas interactúan en levadura o *«in vitro»*

por su estructura y configuración molecular. Sin embargo, al estar las proteínas fuera de su contexto biológico, aunque la técnica indique que se ha producido una interacción, ésta puede no tener sentido biológico en la especie de origen, debido a que dichas proteínas no coinciden en el tiempo –por expresarse en distintos estadios o procesos– o en el espacio –por ejemplo si se expresan en distintos tejidos–. Esto genera falsos positivos, o lo que es lo mismo, pares de proteínas anotadas como interactuantes pero que realmente no lo son en el organismo que se estudia. La elección de una u otra técnica dependerá de la pregunta científica que se quiera responder. A veces ambas se combinan para dar fiabilidad y una visión más completa de las interacciones.

Además del problema de los falsos positivos derivados de las condiciones experimentales en que se determina la interacción, los resultados negativos –proteínas que no interaccionan– no suelen publicarse. La existencia tanto de falsos positivos como de falsos negativos hace que a menudo los datos de interacción no se consideren completamente verosímiles, aunque esto puede paliarse combinando datos positivos que aparezcan en distintas fuentes simultáneamente (Mazandu and Mulder, 2014) y seleccionando conjuntos negativos fiables (Trabuco et al., 2012; Blohm et al., 2014). Se pueden considerar las interacciones como una representación de las relaciones funcionales entre genes: si dos proteínas interactúan físicamente es posible que los genes que las codifican compartan función por estar involucrados en un mismo proceso (Sharan et al., 2007).

1.1.11. Visualización y clasificación de datos

Hasta ahora se han tratado aquí los perfiles fenotípicos como descriptores de los fenómenos que se manifiestan tras el silenciamiento de un gen. Matemáticamente, cada uno de esos perfiles es un vector que puede ser visualizado como un punto en un espacio n-dimensional, siendo n el número de fenotipos total del ensayo. Sin embargo, dado que el número de dimensiones suele ser alto, la visualización de dichos puntos se hace compleja. Es por ello necesario reducir las dimensiones de los parámetros extraídos.

1.1.11.1. Reducción de dimensiones

La asunción general en el reconocimiento de patrones es que dos perfiles próximos en el espacio fenotípico deben también parecerse al observarlos en el mundo real. Así, si partimos de un punto –que representa un perfil– en el espacio fenotípico, las posiciones cercanas deben representar perfiles muy similares. Es decir, se asume que los puntos no están aleatoriamente distribuidos por el espacio fenotípico, sino que forman patrones o nubes de puntos que se asemejan entre ellos. Sin embargo, la dispersión de los puntos en un espacio multidimensional es alta y además aumenta exponencialmente con el número de dimensiones. Este fenómeno es conocido como la *maldición de la dimensionalidad* («*curse of dimensionality*») (Bishop, 2007). Pero este no es el único motivo por el que reducir las dimensiones de los perfiles es imprescindible. Por ejemplo, los humanos tenemos una gran capacidad visual para la detección de patrones, por lo que la observación directa de los puntos en el espacio puede revelar nuevas agrupaciones. Al reducirse el número de dimensiones, se facilita la visualización por ser más compacta y menos redundante.

El objetivo es, por tanto, eliminar dimensiones de los datos sin que se pierda demasiada información del conjunto de datos original. Una forma de alcanzarlo consiste en reducir el número de dimensiones hasta mantener únicamente las más representativas, mediante un proceso conocido como *selección de variables*. Al descartar las variables correlacionadas, se elimina información superflua o redundante, y las relaciones entre las características que permanecen son variables independientes u ortogonales.

Otra opción puede ser calcular un pequeño número de nuevas características que sean una combinación –lineal o no– de las originales. Se distinguen, en este caso, dos tipos de aproximaciones: por un lado, los métodos no lineales, como el escalado multidimensional (*MDS, «MultiDimensional Scaling»*, Definición 26) y los lineales, que conservan el sistema de coordenadas ortogonal, como el análisis de componentes principales (*PCA, «Principal Component Analysis»*, Definición 27) (Jolliffe, 2002).

> *Definición* 26. **Escalado multidimensional (MDS, «MultiDimensional Scaling»)**: método de visualización del grado de similitud de un conjunto de elementos entre los que se ha calculado la distancia. El objetivo es construir una equivalencia con un conjunto de dimensiones más reducidas de forma que las distancias originales se conserven lo máximo posible.

El escalado multidimensional o *MDS* coloca cada punto en un espacio multidimensional de forma que las distancias se conserven lo máximo posible con respecto a la distribución de puntos original. Este grado de correspondencia entre las distancias –la previa y la calculada tras la transformación– se conoce como *estrés de Kruskal*, cuyo valor ha de ser lo menor posible para que la representación sea fiel a la distribución de partida.

> *Definición* 27. **Análisis de componentes principales (PCA, «Principal Component Analysis»)**: método estadístico para la transformación lineal (rotación) de un conjunto de puntos a un nuevo espacio de características no correlacionadas entre sí.

En la aproximación lineal, el análisis de componentes principales transforma las variables para obtener un espacio con menor dimensionalidad. Las dimensiones resultantes se suelen ordenar por la proporción de varianza explicada en los datos, aunque hay muchos otros criterios (Jolliffe, 2002). Aquellas variables que mejor capturan la variabilidad de los datos pasan al nuevo espacio de características que, una vez transformado, será un conjunto de variables ortogonal y sin correlación. La utilidad de analizar las componentes principales no está centrada únicamente en la visualización de datos, sino que también se aplica en la etapa de preprocesamiento para su clasificación.

En el caso de los perfiles fenotípicos, al tratarse de vectores binarios, es decir, formados por cadenas de ceros y unos, es más apropiado usar el *PCA logístico*, que interpreta los datos como

probabilidades de *Bernoulli*. Esta variante reduce el número de dimensiones sin depender del número de vectores al que se aplique, resolviendo más eficientemente el problema mediante una multiplicación matricial (Landgraf et al., 1999).

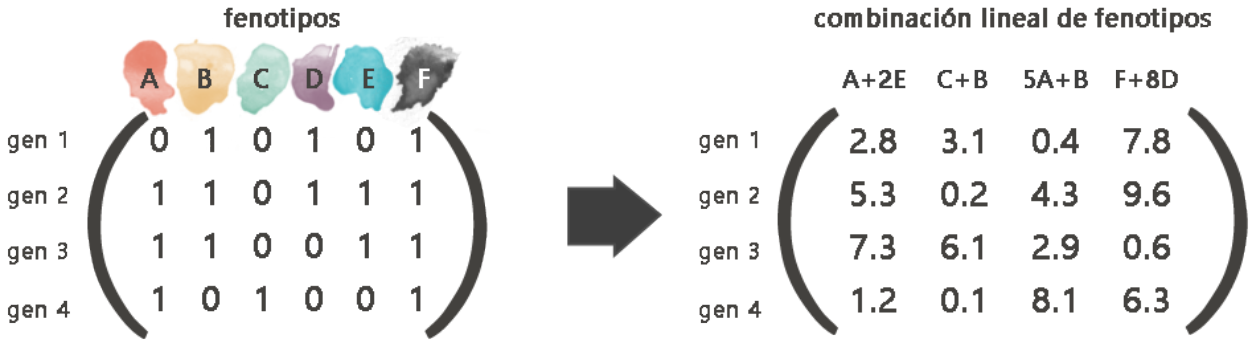


Figura 1.21: **Ejemplo de reducción de una matriz gen-fenotipo usando PCA.** La matriz original (izquierda) representa la asociación entre genes y fenotipos. Al aplicar la reducción de la matriz mediante PCA, las nuevas columnas se transforman en combinaciones lineales de los fenotipos originales y los perfiles fenotípicos de los genes se expresan en ese nuevo espacio.

Una vez reducida la matriz original usando PCA (Figura 1.21), los perfiles fenotípicos que se obtienen no son enteros –ni binarios–, pues representan a los vectores originales en un nuevo espacio donde las variables ahora son combinaciones lineales de los fenotipos de partida. Esto repercute a la hora de calcular similitudes entre los genes en el nuevo espacio: los genes ya no están anotados en términos –sea función o fenotipo– sino en combinaciones lineales de los mismos. Esto dificulta el cálculo de la similitud semántica. En el caso de las métricas aplicadas sobre vectores este problema no se da, a menos que la métrica requiera de vectores de entrada binarios.

1.1.11.2. Algoritmos de agrupamiento

El objetivo de aplicar las métricas descritas en la sección sobre medidas de similitud vectorial (sección 1.1.6) es determinar la similitud entre perfiles. De esta forma, se pueden agrupar los genes en función de su similitud fenotípica. Para ello se usan los algoritmos de clasificación.

De forma análoga a los métodos de visualización, en los algoritmos de aprendizaje automático (Definición 1) también se distinguen dos aproximaciones: lineales y no lineales. Cuando la aproximación es lineal, se asume que las características van a poder distinguirse en un sistema de coordenadas ortogonal. Sin embargo, cuando la relación entre las variables es más compleja, se aplican los métodos no lineales. Este es el caso más habitual para los datos biológicos, pues dada su complejidad intrínseca, los métodos lineales no suelen ajustarse bien. No obstante, generar modelos para la descripción de fenotipos es una tarea empírica y no siempre se sabe con anterioridad qué técnica puede funcionar mejor.

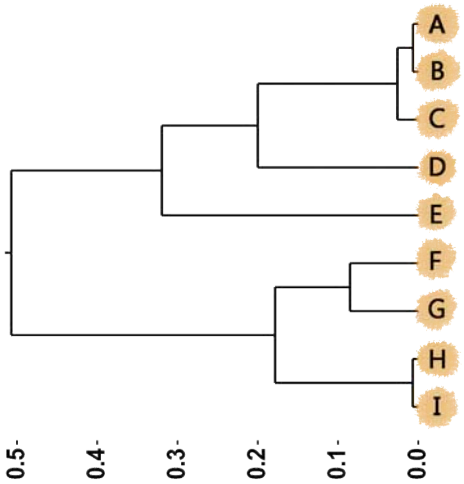


Figura 1.22: **Ejemplo de agrupamiento jerárquico.** En este ejemplo se agrupan distintos elementos (A-I) donde se distinguen dos ramas claramente. Los elementos A y B están muy cercanos (se puede comprobar que la distancia es 0 en la escala inferior); mientras que A y F presentan una distancia cercana a 0.5.

Aunque se dispone de un amplio rango de algoritmos de clasificación aplicable a perfiles, la mayoría de estudios usan el *clustering* o **agrupamiento jerárquico** (Greenacre and Primicerio, 2013), un método de aprendizaje no supervisado (Definición 28) en el que los perfiles se agrupan en una jerarquía que se visualiza en un dendrograma (Figura 1.22). Este método consiste en asignar a cada muestra un grupo o clúster al inicio, para luego ir mezclando los clústeres más cercanos hasta que todas las muestras están en un único grupo que abarca el conjunto

completo de datos. La proximidad entre dos perfiles se mide por la longitud de la rama a la que pertenecen. El algoritmo en sí mismo no define cuales serían los clústeres sino que muestra un dendrograma y es la persona que estudia el fenómeno quien selecciona el punto de corte en el árbol que dará el número de clústeres más adecuado para sus propósitos.

> *Definición* 28. **Agrupamiento no supervisado**: método usado en aprendizaje automático para determinar cómo se organizan los datos sin tener en cuenta un conjunto de entrenamiento previo (para el que se conocería la salida esperada).

Cuando se conoce de forma anticipada el número de grupos que se quiere hacer, el algoritmo más extensamente aplicado es **«K-medias»** o ***«K-means»***. Este algoritmo se inicia fijando unos centroides aleatoriamente en el espacio donde se distribuyen los datos. La idea es encontrar grupos cercanos a esos centroides, para lo que se va corrigiendo la posición de dichos centroides a lo largo de un número dado de iteraciones. Esta aproximación es útil, pero los datos no siempre son clasificables en cuanto a la posición que ocupan en el espacio, sino que se hace necesario proyectar ese espacio en otro de dimensiones mayores donde sí sean separables los puntos. Es esto lo que resuelve el **agrupamiento espectral** o ***«spectral clustering»*** (von Luxburg, 2007), que tiene la capacidad de detectar patrones que otros algoritmos no detectan (Figura 1.23), además de calcularse de manera eficiente mediante álgebra matricial.

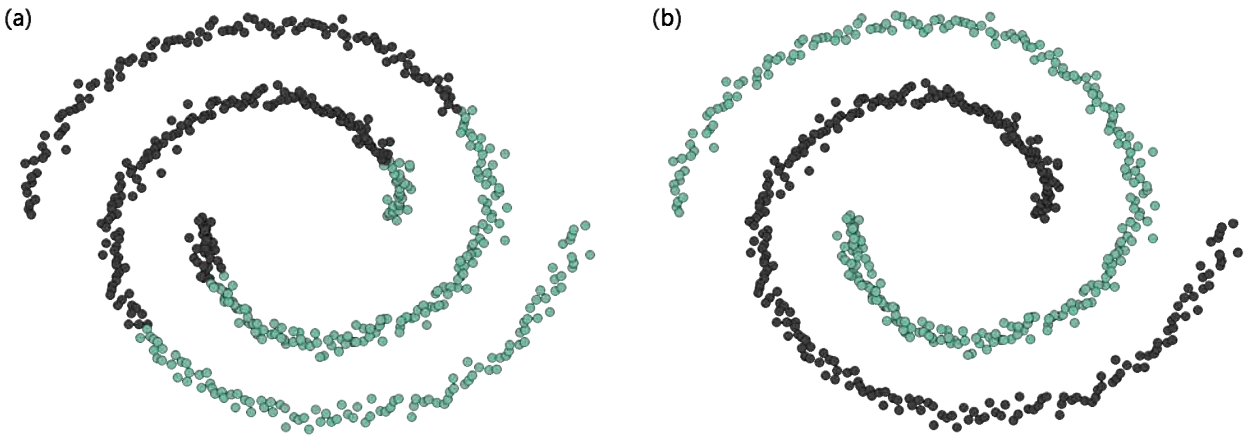


Figura 1.23: **Comparación de «K-means» y «spectral clustering».** Para una misma distribución de puntos en el espacio en forma de espiral, el algoritmo K-means **(a)** crea dos grupos en función de sus centroides, mientras que «spectral clustering» **(b)** separa las dos espirales.

Un punto controvertido en los algoritmos de agrupamiento es la decisión del número de clústeres. En el caso del agrupamiento jerárquico la decisión es del usuario tras la observación del dendrograma. En el algoritmo de las K-medias hay que fijar al inicio el número de centroides, por lo que termina siendo una decisión arbitraria. En este sentido, el agrupamiento espectral proporciona una orientación más fiable calculando los autovalores de la **matriz laplaciana**, que permite extraer propiedades topológicas del grafo. Hay distintas variantes para calcular esta matriz: laplaciana simple, normalizada, generalizada, etc. (Newman, 2010) aunque todas hacen uso de la **matriz de grado** que contiene el número de nodos con el que un nodo se relaciona en la red.

1.1.12. Validación de hipótesis con resultados experimentales

Como se ha visto hasta ahora, los resultados derivados de los experimentos de silenciamiento génico mediante técnicas de RNAi suelen derivar en un extenso listado de genes, lo que complica la interpretación, generación de hipótesis y validación de la misma en un contexto biológico (Figura 1.24).

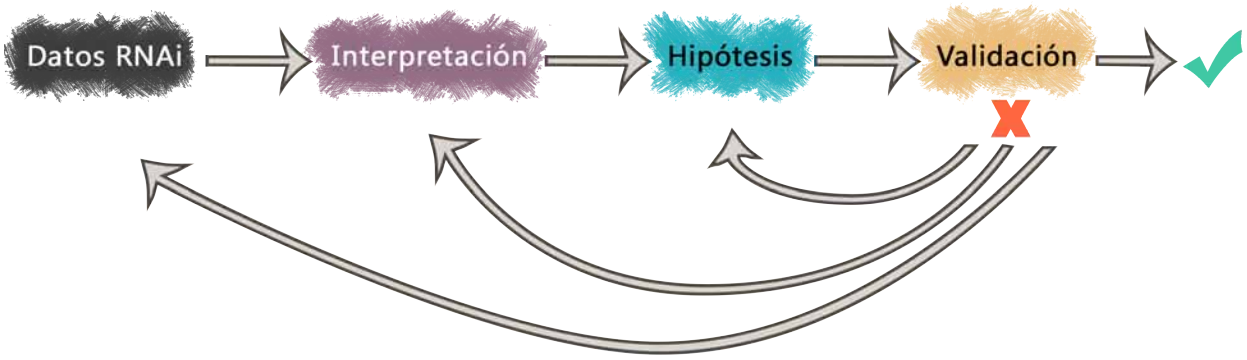


Figura 1.24: **Proceso de validación de una hipótesis.** La interpretación de los datos genera hipótesis falsables que han de ser validadas. Si la validación falla, hay que replantear algunas de las etapas anteriores.

Cuando ya se han recolectado los datos, se interpreta su significado para generar una hipótesis falsable. Lo habitual es agrupar los genes para observar las particularidades por grupos, esto es, ver cómo de similares son entre sí (secciones 1.1.6 y 1.1.8) y qué los distingue del grupo general (esto se conoce como estudios de enriquecimiento). Además de los algoritmos de agrupamiento sugeridos en la sección 1.1.11.2, se pueden considerar también los términos de GO como agrupaciones. Por ejemplo, un conjunto de genes que comparta proceso biológico, función molecular o compartimento celular puede considerarse como una agrupación de acuerdo a un criterio. Esto permite estudiar el enriquecimiento de ciertas funciones.

Una primera exploración de los datos ayuda al planteamiento de la hipótesis de partida o **hipótesis nula**. Generalmente, la hipótesis nula es que las agrupaciones propuestas tras la interpretación de los datos no son correctas. Es decir, lo que refleja una métrica o la organización que sugiere un algoritmo de agrupamiento, no tiene significado biológico.

Para la validación de la hipótesis planteada es necesario, en primer lugar, crear un conjunto positivo y otro negativo. La hipótesis será validada si en el experimento hay más elementos en el conjunto positivo que en el negativo. Este conjunto positivo puede ser una red de interacción

de proteínas (PPI, Definición 25), una función que mida la similitud entre secuencias, una categoría definida por una vía metabólica en KEGG (Ogata et al., 1999), DIP (Salwinski et al., 2004), PANTHER (Mi et al., 2017), etc.

También es necesario definir el conjunto negativo. En el caso de la interacción de proteínas se ha creado recientemente el *«negatome»* (Blohm et al., 2014), donde se definen pares que se ha comprobado que no son interaccionantes. Si no se usa esta categorización, lo habitual es seleccionar aleatoriamente un subconjunto N veces –del mismo tamaño que el conjunto positivo– del grupo de pares de proteínas restantes al quitar el positivo. Se asume que la probabilidad de encontrar un falso negativo en ese conjunto es muy baja.

Una vez definidos los conjuntos positivo y negativo, ya se puede evaluar la capacidad de discriminación de cada medida de similitud. Es decir, la capacidad de una métrica de separar las distribuciones de casos positivos de los negativos. En el caso de las PPIs, se suele evaluar la especificidad y sensibilidad de la métrica mediante una curva ROC (*«Receiving Operating Characteristic»*, Definición 29). Como valor numérico que estima la bondad de la predicción, basándose en la curva ROC, se suele calcular el área bajo la curva (AUC, *«Area Under the Curve»*, Definición 30). Este método es muy común para evaluar las métricas de similitud semántica usando como conjunto positivo las PPIs (KEGG en Guo et al. (2006) y Jain and Bader (2010); DIP en Xu et al. (2008)).

> *Definición* 29. **Curva ROC (*«Receiving Operating Characteristic»*):** La curva ROC permite estimar la calidad de la predicción comparando dos parámetros: razón de verdaderos positivos (o sensibilidad) y razón de verdaderos negativos (1 - especificidad). Se usa para comparar la predicción con respecto a los conjuntos positivo y negativo.

> *Definición* 30. **AUC (*«Area Under the Curve»*):** Medida numérica de la calidad de una predicción basada en el área resultante tras el cálculo de la curva ROC, donde un área de 1 representa el predictor perfecto y 0.5 un rendimiento no superior a lo que se podría haber obtenido por azar.

Otro método que permite evaluar la capacidad discriminatoria de las medidas de similitud es el test estadístico de Mantel (Mantel, 1967). Este procedimiento se basa en la comparación matricial, por lo que tanto el conjunto positivo –interacciones físicas entre proteínas– como la medida de similitud a evaluar deben expresarse como matrices. De esta forma podría medirse cómo de acertada es dicha métrica para reflejar la interacción entre proteínas y por tanto, su similitud funcional.

1.2. Estructura

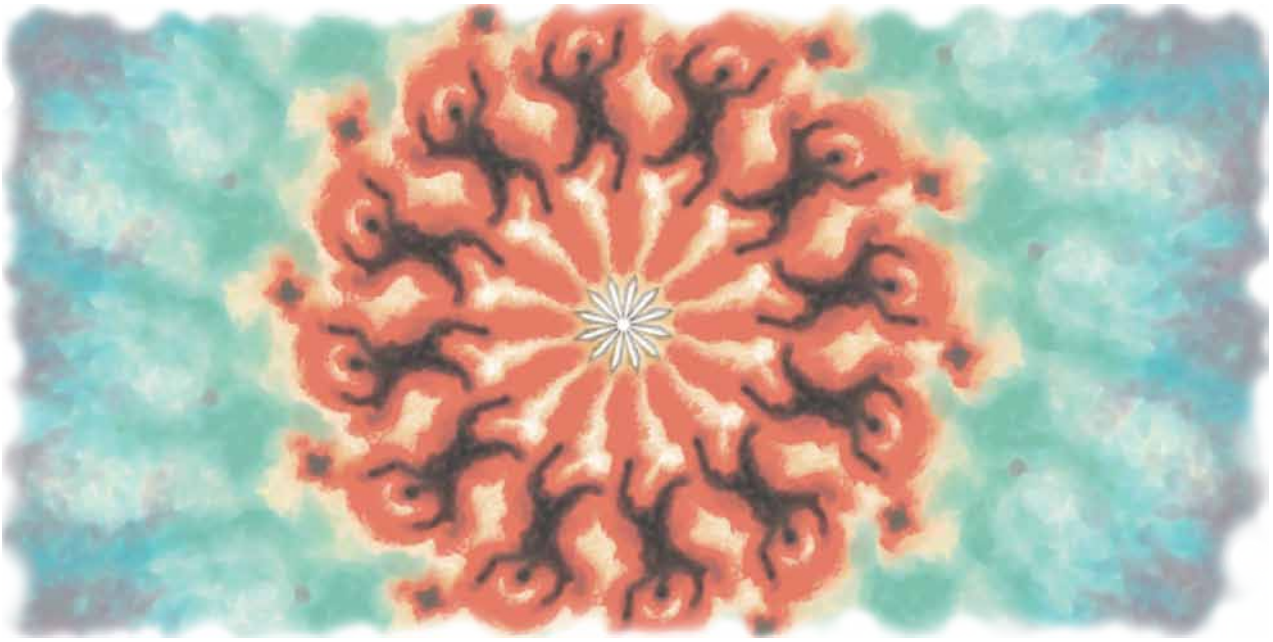
Esta tesis está estructurada del siguiente modo: en el Capítulo 2 se describe brevemente el problema; en el Capítulo 3 se enumeran los objetivos; el Capítulo 4 detalla la metodología seguida para el cumplimiento de los objetivos planteados. A lo largo del Capítulo 5 se describen y discuten los resultados obtenidos, que terminan resumidos en el Capítulo 6, junto con las implicaciones que pueda tener para su desarrollo futuro. Finalmente, en el Capítulo 7 se enumeran las conclusiones finales.

Every explanation is after all an hypothesis.

Ludwig Wittgenstein (1889 – 1951)

2

Descripción del problema. Hipótesis



En el ámbito de la Biología Molecular, uno de los métodos tradicionales que se aplica para asignar funciones a los genes se basa en la inhibición o silenciamiento de los mismos para observar su fenotipo y de ahí, inferir su función. Un principio generalmente aceptado es que la inhibición de genes implicados en un mismo proceso biológico debería mostrar fenotipos parecidos. Este principio se ha extrapolado al análisis sistemático de datos obtenidos por las nuevas técnicas de silenciamento de alto rendimiento, aunque la asociación entre los fenotipos obtenidos y las funciones del gen inhibido no siempre es clara.

Gracias a la ya mencionada *«Microscopía de Sistemas»*, se ha empezado a describir de manera sistemática el espacio fenotípico. Estos experimentos masivos caracterizan al gen mediante un *«perfil fenotípico»* (sección 1.1.4), esto es, una medida multiparamétrica que indica la aparición de un conjunto de fenotipos ante el silenciamiento de un determinado gen. Esta descripción en forma de perfil fenotípico puede parecer en principio más completa por recoger distintas medidas, pero la asociación con la función biológica resulta poco evidente en la mayoría de los casos. Por ello, asociar sistemáticamente una función biológica a un gen en base a su perfil fenotípico es un problema aún no resuelto. La principal consecuencia que se deriva de esto es que la anotación funcional en base al fenotipo suele hacerse de forma manual, con la inversión de tiempo que ello conlleva y el riesgo en cuanto a la propensión a introducir errores y sesgos propios de la persona que anota en cada caso (ver la sección 1.1.3.3).

Además de la problemática derivada de la anotación manual, evaluar la similitud entre dos fenotipos tampoco resulta trivial, particularmente al integrar datos de distintos experimentos. La dificultad reside en encontrar tanto la métrica adecuada como un valor umbral de similitud entre fenotipos que permita inferir que éstos comparten una misma función. Por ejemplo, se pueden comparar fácilmente las secuencias de dos genes o proteínas a través de algoritmos de alineamiento, pero no ocurre lo mismo en el caso de la comparación de vectores con parámetros fenotípicos o funcionales. Es por ello que a día de hoy, los experimentos de silenciamiento de alto rendimiento a partir de técnicas de siRNA no suelen ser usados para anotar genes en términos de *Gene Ontology*.

Un ensayo experimental es el método más fiable para asignar una función a un gen. Sin embargo, en la era de la secuenciación de alto rendimiento existen serias dificultades para llevar a cabo experimentos que determinen la función de cada uno de los productos de genes. Por tanto, para ganar profundidad en el conocimiento de la actividad de estas moléculas y guiar adecuadamente los experimentos, debemos desarrollar algoritmos y métodos computacionales que nos permitan anotar funcionalmente de forma precisa a partir de los datos fenotípicos generados por las aproximaciones experimentales de alto rendimiento. Hay multitud de ensayos experimentales que se centran en la predicción funcional de genes cuyos estudios sostienen que los genes que se predicen que están relacionados funcionalmente, también presentan fenotipos similares (Lee et al., 2008; Qi et al., 2008; Hu et al., 2009; Rojas et al., 2012; Hériché et al., 2014).

Partiendo de la abundancia y heterogeneidad de los experimentos, la automatización de los análisis se hace fundamental: los fenotipos que son similares se asocian a procesos celulares también similares. Si este proceso se automatiza, entonces el número de anotaciones funcionales procedentes de experimentos puede aumentar, contribuyendo a la mejora del conocimiento biológico y clínico.

Como ya se ha indicado en la Introducción de esta Tesis Doctoral, la organización de la información es un punto importante para la automatización del proceso. Se han hecho algunos intentos para organizar los fenotipos observados a nivel celular. Entre estas aproximaciones están CPO (*Cellular Phenotype Ontology*, Hoehndorf et al. (2012)) y CMPO (*Cellular Microscopy Phenotype Ontology*, Jupp et al. (2016)), que organizan los fenotipos de forma consistente. El hecho de que existan estas descripciones fenotípicas basadas en una estructura y no como texto plano, hace posible la evaluación de la similitud entre fenotipos de forma automática.

Para poder llevar a cabo la conversión de anotación de un perfil fenotípico a anotación funcional, necesitamos entender cómo están relacionados genéticamente los fenotipos y las funciones. La mayoría de experimentos genera un listado de genes involucrados en un mismo proceso biológico atendiendo a anotaciones en GO. Esto parece indicar que la similitud fenotípica ha de presentar correlación con la similitud funcional.

2.1. Hipótesis

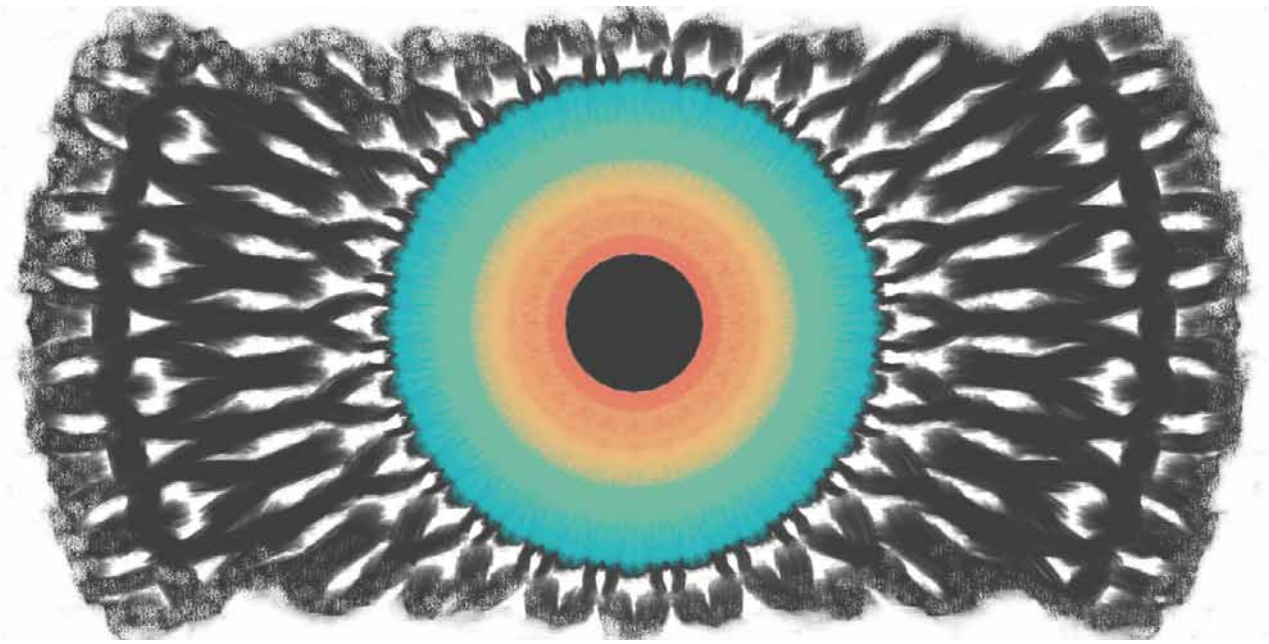
Por todo lo anteriormente expuesto, la hipótesis principal de esta Tesis Doctoral es que los fenotipos que los genes manifiestan pueden describirse mediante un perfil descriptivo de la función o funciones que dichos genes desempeñan y que esto es posible a través del estudio de la relación entre las anotaciones fenotípicas y funcionales de estos genes.

A goal without a plan is just a wish.

Antoine de Saint-Exupéry (1900 – 1944)

3

Objetivos



Haciendo uso de las estructuras ontológicas así como de las herramientas informáticas para la interpretación de datos biológicos, en esta Tesis Doctoral se establecen los siguientes objetivos para verificar la hipótesis planteada:

❶ Comparación de distintas medidas de similitud fenotípica y selección de la más adecuada para representar la relación funcional entre pares de genes descritos a través de sus perfiles fenotípicos.

❷ Estudio de la relación entre las similitudes funcional y fenotípica de pares de genes entre distintos dominios ontológicos para observar cómo las anotaciones fenotípicas pueden derivarse de las anotaciones funcionales y viceversa.

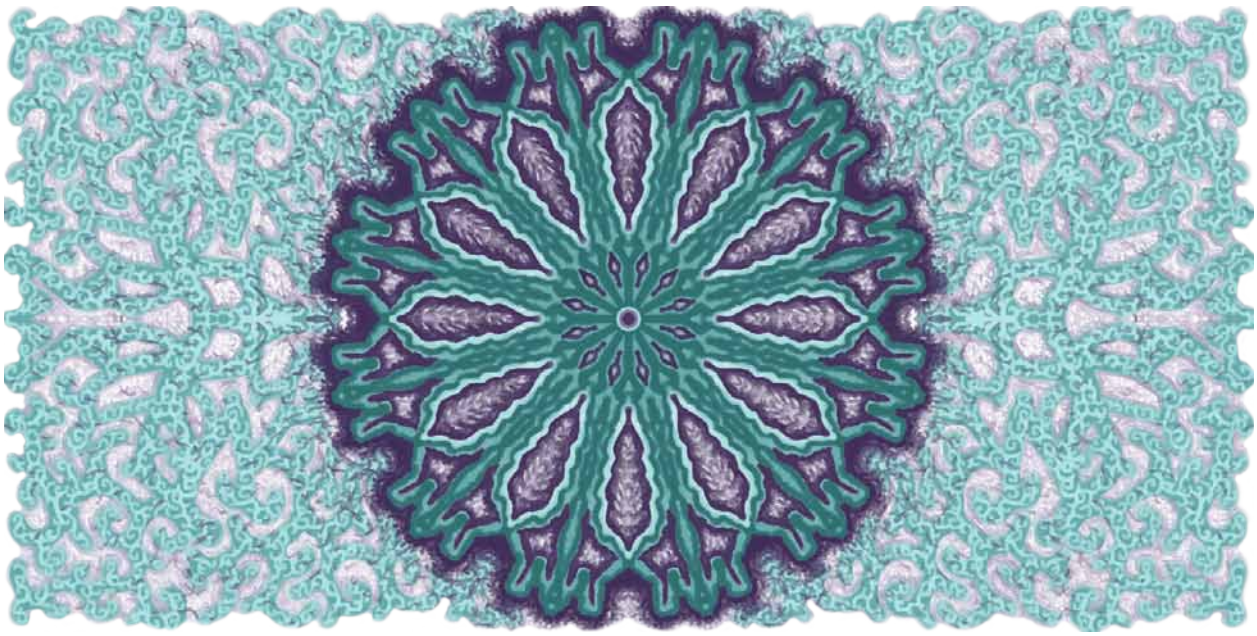
❸ Exploración de una organización alternativa de las estructuras ontológicas que relacione de forma biunívoca los espacios fenotípico y funcional.

When you can measure what you are speaking about, and express it in numbers, you know something about it.

Lord Kelvin (1824 – 1907)

4

Material y Métodos



En este capítulo se detalla la metodología seguida durante el desarrollo de esta Tesis Doctoral así como los recursos utilizados. Toda esta información se encuentra tanto en bases de datos biológicas como en conjuntos de datos experimentales.

4.1. Anotaciones funcionales y fenotípicas

En este trabajo se han usado principalmente dos ontologías formales con sus respectivas anotaciones: *Gene Ontology* (GO) (Ashburner et al. (2000)) y *Cellular Microscopy Phenotype Ontology* (CMPO)[1] (Jupp et al., 2016). Se ha seleccionado la rama de procesos celulares, cuyo término raíz es *«cellular process»* (*GO:0009987*). La selección de este término viene dada por ser la que mejor se ajusta a los fenotipos –centrados en procesos celulares– estudiados en los experimentos de silenciamiento génico.

La jerarquía de términos GO se extrajo del fichero OBO (véase la Sección 1.1.7) con versión del 26/09/2015. Las anotaciones se descargaron de la página web de GO[2] con fecha de validación del 16/09/2015. En total, 17232 genes humanos están anotados en 11641 términos GO distintos y el número de relaciones gen-término GO asciende a 230877. Tras eliminar las anotaciones inferidas electrónicamente (IEA), se obtuvieron 139243 relaciones entre 12550 genes y 9200 términos GO distintos.

En cuanto a CMPO, las anotaciones se extrajeron de la base de datos *«Cellular Phenotype Database»*[3] (Kirsanova et al. (2015)). En total, se obtuvieron 13866 relaciones entre 8109 genes y 36 fenotipos.

La ventaja que aporta contar con la ontología frente a un vocabulario no estructurado de fenotipos es que formaliza las relaciones jerárquicas entre los fenotipos (más detalles en la sección 1.1.7).

[1] http://www.ebi.ac.uk/cmpo
[2] http://geneontology.org/page/download-annotations
[3] http://www.ebi.ac.uk/fg/sym

4.2. Fuentes experimentales de asociación entre fenotipos y genes

Los fenotipos observados como consecuencia de la pérdida de función de genes se han obtenido de distintas fuentes experimentales. Es importante destacar que todos los experimentos han sido llevados a cabo en células humanas, aunque en distintas líneas celulares (véase la Tabla 4.1 para más detalles).

Se han usado en este trabajo dos tipos de experimentos: por un lado, aquellos realizados sobre el genoma completo; y por otro, aquellos en los que previamente se ha preseleccionado un conjunto de genes a silenciar.

Entre las fuentes de datos experimentales sobre el genoma completo se encuentran: *CellMorph* (Fuchs et al., 2010), que se centra en la detección de los cambios morfológicos que provocan en las células la inhibición de cada uno de los genes que conforman el genoma humano; *MitoCheck* (Neumann et al., 2010), que estudia los genes que pueden estar involucrados en procesos relacionados con la mitosis y con el movimiento celular; y *EMBL secretion* (Simpson et al., 2012), que analiza el transporte desde el retículo endoplasmático hasta la membrana para la secreción de la proteína de carga *ts045G*. De la base de datos *GenomeRNAi* (Schmidt et al., 2013) se han utilizado los datos correspondientes a dos experimentos que también se han llevado a cabo sobre el genoma completo. Estos experimentos son: *GR00053* (Paulsen et al., 2009) y *GR00290* (Balestra et al., 2013), que analizan el daño del DNA en células *HeLa* y la regulación de la formación del centriolo, respectivamente.

Por otro lado, se han usado datos de dos experimentos en los que se silencian subconjuntos de genes específicos para el estudio de la respuesta celular ante radiación ionizante (*Copenhagen DNA damage Ubiquitin*, Moudry et al. (2012)) y la condensación cromosómica (*EMBL chromosome condensation*, Hériché et al. (2014)).

En total, agrupando todos los datos de los distintos experimentos se obtiene un conjunto de 36 fenotipos únicos no solapantes que aparecen descritos textualmente en la tercera columna de la Tabla 4.1.

Experimento	Descripción	Fenotipos	IDs en CMPO
CellMorph (Fuchs et al., 2010)	Análisis tras la inhibición mediante RNAi del genoma completo que examina los cambios morfológicos de células *HeLa* individuales dentro de poblaciones.	decreased cell number cell with projections elongated cell more lamellipodia cells increased number of actin filament round cell increased cell size decreased cell size bright nuclei metaphase arrested increased cell size in population	CMPO:0000052 CMPO:0000071 CMPO:0000077 CMPO:0000083 CMPO:0000105 CMPO:0000118 CMPO:0000128 CMPO:0000129 CMPO:0000154 CMPO:0000305 CMPO:0000340
MitoCheck (Neumann et al., 2010)	Análisis tras la inhibición mediante RNAi del genoma completo de la segregación cromosómica en células *HeLa*. También se estudian genes involucrados en otros procesos como el movimiento celular.	cell death increased nucleus size graped micronucleus abnormal nucleus shape mitosis delayed binuclear cell absence of mitotic chromosome decondensation increased cell movement speed increased cell movement distance proliferating cells metaphase delayed abnormal chromosome segregation prometaphase delayed increased variability of nuclear shape in population mitotic metaphase plate congression	CMPO:0000030 CMPO:0000140 CMPO:0000156 CMPO:0000157 CMPO:0000202 CMPO:0000213 CMPO:0000216 CMPO:0000236 CMPO:0000237 CMPO:0000241 CMPO:0000307 CMPO:0000326 CMPO:0000344 CMPO:0000345 CMPO:0000348
EMBL secretion (Simpson et al., 2012)	Análisis tras la inhibición mediante RNAi del genoma completo para la interferencia con el transporte RE-membrana plasmática para la secreción de la proteína de carga *ts045G* en células *HeLa*.	increased rate of protein secretion mild decrease in rate of protein secretion strong decrease in rate of protein secretion decreased rate of intracellular protein transport	CMPO:0000246 CMPO:0000318 CMPO:0000319 CMPO:0000346
GR00053 (Paulsen et al., 2009)	Análisis tras la inhibición mediante RNAi del genoma completo para genes involucrados en la respuesta al daño del DNA en células *HeLa*.	increased number of site of double-strand break	CMPO:0000182
GR00290 (Balestra et al., 2013)	Análisis tras la inhibición mediante RNAi del genoma completo de genes reguladores de la formación del centriolo en células *HeLa*.	increased centriole replication decreased centriole replication	CMPO:0000361 CMPO:0000362
Copenhagen DNA damage Ubiquitin (Moudry et al., 2012)	Análisis tras la inhibición mediante RNAi de más de 1300 genes involucrados en el sistema de la ubiquitina-proteosoma o que codifican proteínas con dedos de zinc buscando moduladores de respuestas celulares a la radiación ionizante en células *HeLa* y *U2OS*.	decreased number of site of double-strand break	CMPO:0000181
EMBL chromosome condensation (Hériché et al., 2014)	Análisis tras la inhibición mediante RNAi de 100 genes seleccionados usando métodos bioinformáticos para analizar los cambios en la duración de la profase mitótica en células *HeLa*.	increased duration of mitotic prophase decreased duration of mitotic prophase	CMPO:0000328 CMPO:0000329

TABLA 4.1: **Conjunto de 36 fenotipos obtenidos de los experimentos de siRNA.** En la primera columna se encuentra el nombre y la cita del estudio experimental, en la segunda una pequeña descripción del mismo, en la tercera los fenotipos que se identifican en cada estudio y en la última columna, el identificador en la ontología CMPO de cada uno de los fenotipos.

Las funciones celulares que abarcan estos estudios son muy diversas, aunque cada uno pone el foco en un tipo de fenómeno. En líneas generales, el conjunto incluye: proliferación celular, muerte celular, mitosis, secreción de proteínas, daño en el DNA y formación del centriolo. Sin embargo, a veces en los experimentos también se anotan fenotipos que no están directamente relacionados con la función biológica en estudio, como es el caso de *MitoCheck*, por lo que pueden aparecer fenotipos que no sean clasificables estrictamente dentro de estas categorías. Una representación gráfica de la frecuencia de aparición de todos los fenotipos obtenidos puede verse en la Figura 4.1.

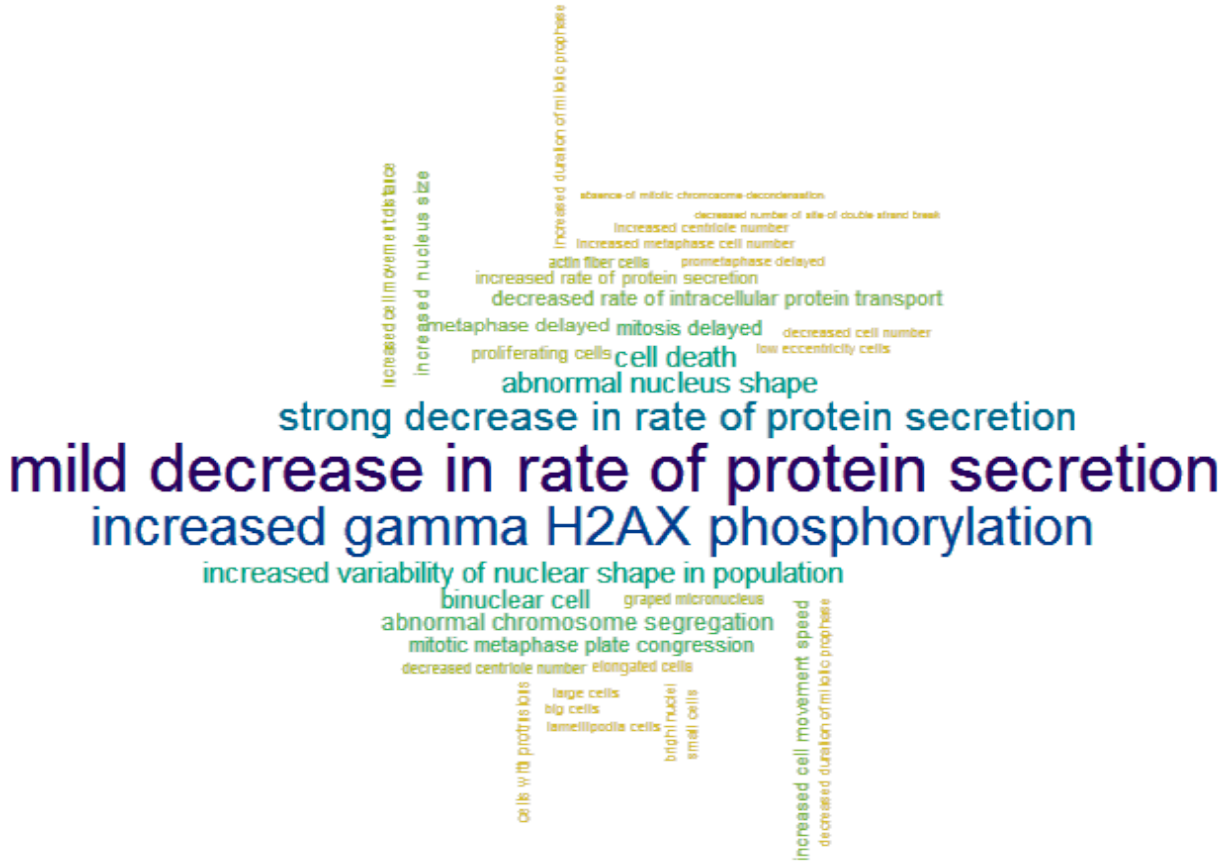


Figura 4.1: **Nube de palabras representando la frecuencia de los fenotipos.** Representación de los 36 fenotipos con tamaños proporcionales a la cantidad de genes que tienen anotados.

Entre todos los experimentos suman un total de 4198 genes asociados a fenotipos –representados mediante sus identificadores *Entrez*–, donde la mayoría de ellos se ha silenciado en más de un experimento. En la Figura 4.2, se puede observar una representación visual de la frecuencia de dichos genes en cuanto al número de fenotipos que presentan.

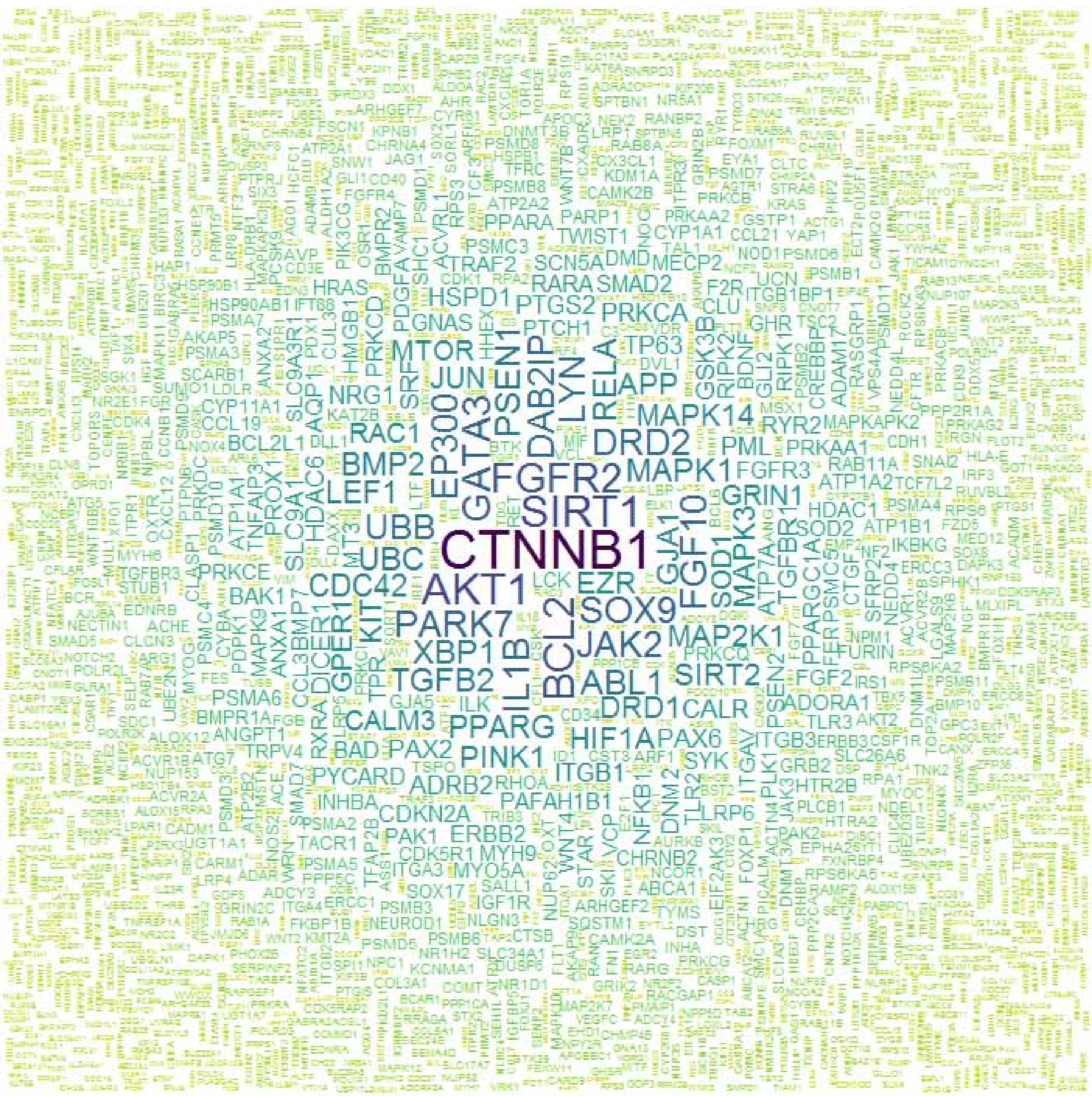


Figura 4.2: **Nube de palabras representando la frecuencia de asociación de los genes a los fenotipos.** El tamaño del identificador en formato *gene symbol* del gen indica la frecuencia de dicho gen en las asociaciones gen-fenotipo.

4.2.1. Matriz integrada de genes y fenotipos

Las relaciones entre los genes y los fenotipos de diferentes ensayos se han integrado en una matriz binaria formada por todos los perfiles, donde el valor 1 representa la presencia del fenotipo y el valor 0, la ausencia del mismo. Un ejemplo de la representación de la matriz puede encontrarse en la Tabla 4.2.

IDENTIFICADOR *Entrez* (**gene symbol**)	**decreased cell number** (CMPO:0000052)	**cell with projections** (CMPO:0000071)	…	**mitotic metaphase plate congression** (CMPO:0000348)
57147 (**SCYL3**)	1	0	…	0
2268 (**FGR**)	1	0	…	1
22875 (**ENPP4**)	0	1	…	0
…	…	…	…	…
5439 (**POLR2J**)	1	0	…	1

TABLA 4.2: **Matriz binaria para las asociaciones gen-fenotipo**. La presencia o ausencia de un fenotipo (columna) tras la inhibición de un gen (fila) se representa con los valores 1 y 0, respectivamente.

En general, el número de fenotipos en que se anota un gen es bajo, lo que hace la matriz dispersa, o lo que es lo mismo, una matriz con un alto porcentaje de ceros (más información sobre las características de la matriz resultante en la Tabla 4.3).

Descripción	**Cantidad**
Número de genes	4198
Número medio de genes por fenotipo	209.28
Máximo número de genes por fenotipo	1469
Mínimo número de genes por fenotipo	3
Número de fenotipos	36
Número medio de fenotipos por gen	1.8
Máximo número de fenotipos por gen	15
Mínimo número de fenotipos por gen	1
Número de perfiles fenotípicos distintos	616
Porcentaje de perfiles que comparten más de un fenotipo	99.95

TABLA 4.3: **Estadísticas sobre la matriz binaria gen-fenotipo**. Descripción de algunos parámetros sobre la matriz que contiene los perfiles fenotípicos de los genes.

Dado que hay experimentos que no incluyen el genoma completo en su estudio, hay genes para los que el valor 0 no representa la ausencia del fenotipo, sino que al no haber sido seleccionado el gen como diana para su silenciamiento en un experimento dado, no es posible obtener un resultado fenotípico asociado. Estos genes que no han sido silenciados y por tanto se desconoce su fenotipo, podrían introducir un sesgo en el estudio. Para caracterizar los efectos que esos datos ausentes pueden tener en los resultados, una aproximación consiste en reemplazar una proporción aleatoria de unos por ceros en la matriz. De esta forma, se puede observar si los resultados son robustos ante valores incrementales de dispersión (en este caso, 5 %, 10 %, 20 % y 30 %).

Como el objetivo es explorar el modo en que las anotaciones fenotípicas están relacionadas con los términos de la rama *«cellular process»* de GO, aquí se usan las anotaciones fenotípicas tal y como se han descrito en cada uno de los artículos. Es importante destacar que ninguno de estos experimentos se ha empleado previamente para anotar genes en GO, como puede comprobarse por el hecho de que ninguno de los artículos citados aparece como fuente de información en GO, a pesar de haber estado los datos disponibles durante varios años. Por tanto, se puede considerar que las anotaciones de la ontología GO son independientes del conjunto de genes integrados en la matriz.

4.2.2. Reducción de dimensiones de la matriz integrada de genes y fenotipos

Para la reducción de dimensiones (sección 1.1.11.1) se aplicó la función `logisticPCA()` del paquete `logisticPCA`, extrayendo las 10 principales componentes de la matriz. En este nuevo espacio se calcularon las similitudes *Euclídea*, *correlación* y *coseno*.

4.3. Medidas de similitud fenotípica

Las medidas de similitud usadas en este estudio se pueden clasificar en vectoriales –véase en la Sección 1.1.6– aplicadas sobre los perfiles fenotípicos (Tabla 4.4) y en aquellas que hacen uso de una estructura ontológica para calcular la similitud (Tabla 4.5), explicadas en la sección 1.1.8.

Nombre	**Fórmula**
Similitud Euclídea	$s^2(g_1,g_2) = \frac{1}{1+(x_{g1}-x_{g2})(x_{g1}-x_{g2})'}$
Similitud correlación	$s(g_1,g_2) = \frac{(x_{g1}-\overline{x}_{g1})(x_{g2}-\overline{x}_{g2})'}{\sqrt{(x_{g1}-\overline{x}_{g1})(x_{g1}-\overline{x}_{g1})'}\sqrt{(x_{g2}-\overline{x}_{g2})(x_{g2}-\overline{x}_{g2})'}}$ donde $\overline{x}_{g1} = \frac{1}{n}\sum_{p\in P} x^p_{g1}$ y $\overline{x}_{g2} = \frac{1}{n}\sum_{p\in P} x^p_{g2}$
Similitud de cosenos	$s(g_1,g_2) = \frac{x_{g1}x'_{g2}}{\sqrt{x'_{g1}x_{g1}}\sqrt{x'_{g2}x_{g2}}}$
Similitud de Hamming	$s(g_1,g_2) = \frac{x^p_{g1}=x^p_{g2}}{n}$
Similitud de Jaccard	$s(g_1,g_2) = 1 - \frac{x^p_{g1}\neq x^p_{g2}}{(x^p_{g1}\neq 0)\vee(x^p_{g2}\neq 0)}$
Similitud de Cohen's kappa	$s(g_1,g_2) = \frac{p_0-p_c}{1-p_c}$ donde: - p_0 es la proporción de términos comunes en los perfiles de g_1 y g_2, y - p_c es la proporción de términos comunes en los perfiles de g_1 y g_2 esperadas por azar.
Similitud TF-IDF	$s(g_1,g_2) = \text{máx}_{p\in P}\{x^p_{g1}x^p_{g2}IDF(p)\}$ donde $IDF(p) = \log\frac{n_G}{1+\sum_{g\in G}x^p_g}$

TABLA 4.4: **Medidas de similitud entre perfiles fenotípicos.** G es el conjunto de n_G genes y P es el conjunto de n_P fenotipos. x_g denota el perfil fenotípico del gen g con $x^p_g = 1$ si g muestra el fenotipo p, $x^p_g = 0$ en otro caso.

Las métricas *Euclídea* y *correlación* se calcularon usando el paquete `stats` de `R`. Para el *coseno* se usó `lsa` (Wild, 2015) y para *Jaccard*, el paquete `prabclus` (Hennig and Hausdorf, 2015). También se implementaron en `R` el resto de métricas: *Hamming*, *Cohen's kappa* y *TF-IDF* (Robertson, 2004). Con la intención de aprovechar la estructura ontológica de los fenotipos, también se calcularon las métricas de similitud semántica de la Tabla 4.5 usando el paquete `dnet` (Fang and Gough, 2014), en `R`.

Nombre	**Fórmula**
Similitud semántica de Resnik (Resnik, 1995)	$s(t_1, t_2) = IC(t_{MICA})$ donde: - el ancestro común más informativo (MICA) es $t_{MICA} = argmax_{t \in S(t_1,t_2)} IC(t)$, - el contenido informativo (IC) de un término t es $IC(t) = -log(p(t))$, - la probabilidad de un término t es $p(t) = \frac{anotaciones(t)}{anotacionesTotales}$, y - $A(t_1, t_2)$ es el conjunto de ancestros comunes de t_1 y t_2.
Similitud semántica de Jiang (Jiang and Conrath, 1997)	$s(t_1, t_2) = 1 + 2 \cdot IC(t_{MICA}) - (IC(t_1) + IC(t_2))$
Similitud semántica de Lin (Lin, 1998)	$s(t_1, t_2) = \frac{2 \cdot IC(t_{MICA})}{IC(t_1) + IC(t_2)}$
Similitud semántica de Schlicker (Schlicker et al., 2006)	$s(t_1, t_2) = \frac{2 \cdot IC(t_{MICA})}{IC(t_1) + IC(t_2)} \cdot (1 - p(t_{MICA}))$
Similitud semántica de Pesquita (Pesquita et al., 2007)	$s(t_1, t_2) = \frac{\sum_{t \in A(t_1,t_2)} IC(t)}{\sum_{t \in P(t_1,t_2)} IC(t)}$ donde: - $P(t_1, t_2)$ es el conjunto de ancestros comunes de t_1 ó t_2.

TABLA 4.5: **Medidas de similitud semántica en ontologías.** Todas las métricas de la tabla se basan en el contenido informativo de los términos que se deriva de las anotaciones.

4.4. Comparación entre medidas de similitud fenotípica

Para evaluar las potenciales relaciones de dependencia entre las distintas medidas de similitud fenotípica se calcularon las similitudes entre pares de genes para cada métrica. Con estos valores, se calculó el grado de relación o similitud entre las métricas de acuerdo al *coeficiente de correlación de Pearson* (PCC). Una vez obtenidas las similitudes entre métricas, éstas se organizaron mediante un agrupamiento o *clustering* jerárquico (Sección 1.1.11.2) usando el método *average-linkage* de la función `hclust` de `R`. Puesto que PCC es una medida de similitud, en el dendrograma se usó $(1 - PCC)$ como medida de distancia.

4.4.1. Capacidad de las métricas de similitud fenotípica para distinguir entre pares interactuantes y no interactuantes

Además de estudiar cómo las métricas se relacionan entre ellas, también se ordenaron por la capacidad de detectar la relación funcional entre los genes. Para ello, se usaron redes de interacción de proteínas como una representación de las relaciones funcionales entre genes.

En primer lugar, se ordenaron las medidas de similitud según el área bajo la curva ROC (AUC). En este contexto, la AUC se puede interpretar como la capacidad de cada métrica para distinguir los pares interactuantes (verdaderos positivos) de los no interactuantes (verdaderos negativos).

Como conjunto positivo en la evaluación se utilizaron las interacciones físicas de `Intact` (Orchard et al., 2014), `MIPS` (Pagel et al., 2005), `DIP` (Salwinski et al., 2004) y `BIOGRID` (Stark et al., 2006). De todas las fuentes, se filtraron sólo aquellos pares identificados por al menos dos métodos experimentales distintos o bien por ser interacciones curadas en `Reactome` (Milacic et al. (2012), Fabregat et al. (2016)), resultando en total 27689 relaciones entre genes.

Como interacciones negativas, se usó el conjunto curado de interacciones en el `MIPS Negatome` (Blohm et al., 2014) y Trabuco et al (Trabuco et al., 2012), que conforman un conjunto de 895790 relaciones negativas (probablemente no existentes) entre pares de genes. Para calcular el AUC se usó el paquete `pROC` de `R` (Robin et al., 2011).

4.4.2. Test de Mantel

Una segunda alternativa es la comparación de matrices usando el test estadístico de Mantel (Mantel, 1967). En este caso, la comparación entre una matriz de una métrica de similitud fenotípica y la matriz de *iRef index* se calculó con la función `mantel` del paquete `vegan`.

4.4.3. Capacidad de las métricas de similitud fenotípica para detectar interacción física entre genes

Otra forma posible de comparar el rendimiento de las distintas métricas consiste en calcular, para cada una de ellas, el número de genes cuyo par más similar fenotípicamente es además un par interactuante en una red de interacción extraída de *iRef index* (versión 14.0, 07/04/2015) (Razick et al., 2008). La red de interacciones consta de 22921 genes conectados a través de 220926 relaciones. En esta aproximación, el método consiste en identificar, para cada métrica y gen, el vecino más cercano (seleccionando al azar un par en caso de empate). Sólo en el caso de que los dos genes, además de ser cercanos fenotípicamente, sean interactuantes en *iRef index*, se considerará para la puntuación de cada métrica. De esta forma, se esperaría que el conjunto de pares con similitudes fenotípicas más altas estuviese enriquecido en pares que interaccionan físicamente.

Para evaluar la significancia estadística de dicha puntuación, se calculó la probabilidad de tener el mismo o mejor valor que una selección aleatoria calculada a partir de la distribución hipergeométrica. La función de `R` usada para este cálculo es *phyper()*.

Partiendo de los 4198 genes asociados a fenotipos utilizados en este estudio, habría un total de $\frac{4198*(4198-1)}{2} = 8809503$ interacciones no direccionales posibles, de las cuales 29649 aparecen como pares interactuantes en *iRef index*. Para una métrica de similitud dada, se seleccionaron las 4198 interacciones con mayor similitud fenotípica (una por gen).

Por tanto, la probabilidad, p, de presentar un número de aciertos mayor a x seleccionando aleatoriamente las interacciones viene dada por:

H_0 (Hipótesis nula): No hay asociación positiva entre la ocurrencia de interacciones en la métrica y las interacciones en *iRef index*.

X = variable aleatoria que describe el número de elementos interactuantes entre los 4198 genes.

k = 4198 genes.

m = 8809503 interacciones posibles dado un conjunto de 4198 genes.

i = 29649 pares interactuantes en *iRef index*.

$ni = m - i$ = 8809503 − 29649 = 8779854 pares no interactuantes en *iRef index*.

$$p(X \leq x | H_0) = \frac{\binom{i}{x}\binom{ni}{k-x}}{\binom{i+ni}{k}} \tag{4.1}$$

En *R*:

$$p = 1 - phyper(x - 1, 29649, 8809503 - 29649, 4198) \tag{4.2}$$

4.5. Agrupamiento guiado por las anotaciones de los genes en las ontologías GO y CMPO

Siguiendo la aproximación abordada por Glass y Girvan (Glass and Girvan, 2015), se construyó un grafo bipartito (Sección 1.1.10) con las funciones en GO y los fenotipos en CMPO conectados a través de los genes anotados en ambos dominios (Figura 4.3a). Proyectando el grafo bipartito GO-CMPO se obtienen dos posibles redes: un grafo formado únicamente por funciones de GO (Figura 4.3b) y otro formado por fenotipos de CMPO (Figura 4.3c), donde el enlace entre dos términos existe sólo en el caso en que compartan al menos un gen. El peso de la arista viene dado por el número de genes que comparten los dos términos (Figura 4.3b y c).

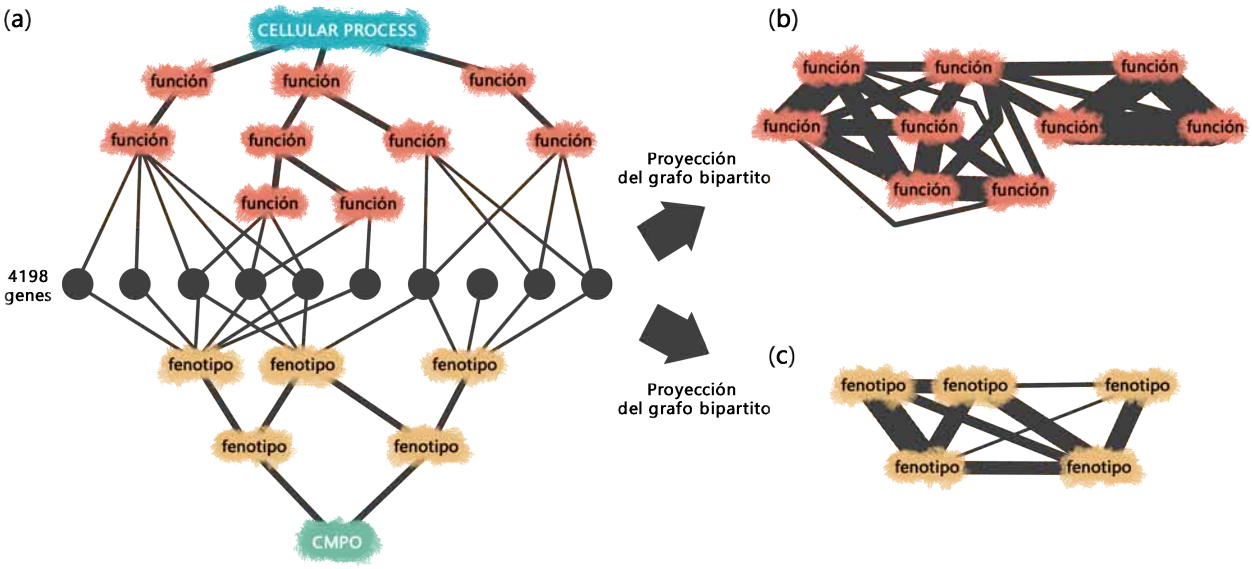


Figura 4.3: **Proyección del grafo bipartito GO-CMPO. (a)** Representación de la rama *«cellular process»* de GO con distintos términos funcionales (naranja) y sus respectivas anotaciones (círculos en gris). Análogamente para la ontología CMPO, los fenotipos (amarillo) tienen anotaciones (gris). **(b)** Red unipartita resultante de la proyección del grafo donde las funciones de GO están conectadas si comparten al menos un gen. La anchura del trazo representa el número de genes compartidos. **(c)** Red unipartita resultante de la proyección del grafo donde los fenotipos se conectan en función del número de genes compartidos, representados por la anchura de la arista.

Los términos que ocupan posiciones más altas en la ontología heredan las anotaciones de sus sucesores, por lo que dichos términos van a tener un grado de conexión muy alto. Para compensar el sesgo, se normalizaron los pesos por la unión de los genes anotados en ambos términos. Después, se agruparon los términos de ambos grafos mediante agrupamiento espectral (Sección 1.1.11.2). De esta forma, se obtuvieron módulos funcionales –términos GO– y fenotípicos –términos CMPO– agrupados por los genes que comparten.

Como está indicado en Glass y Girvan (Glass and Girvan, 2015), el número de clústeres resultante del agrupamiento elegido depende de los distintos niveles de especificidad. Una orientación del número de agrupaciones estimado se extrae del cálculo de los autovalores de la matriz laplaciana de la red. Siguiendo dicha orientación, en este trabajo se fijó el número de clústeres en 140 para el caso de GO y 13 para CMPO. En la selección del número de clústeres en GO, se tuvo en cuenta que el mayor número de agrupaciones tuviera fenotipos vinculados. Para ello, se agruparon primero los nodos de la red en 100 clústeres y luego se seleccionaron los dos de mayor tamaño, a los que se les repitió el proceso para obtener 33 y 9 clústeres, respectivamente. En el caso de CMPO, el número de clústeres de términos CMPO se eligió para producir una distribución de tamaños razonable que minimizara el número de agrupaciones con un único término. Al incrementar este valor, aumentó también el número de clústeres conteniendo sólo un término.

4.6. Ajuste del p-valor por contrastes múltiples

Los p-valores se han ajustado usando la función `p.adjust()` de `R` con el método de Benjamini y Hochberg. Esta función devuelve los p-valores ajustados por dicho método, los cuales son posteriormente comparados con la proporción de falsos positivos (*False Discovery Rate*, FDR) elegida. Este método es el más habitual para contrastes múltiples, siendo menos estricto que la correción de Bonferroni pero proporcionando un buen equilibrio entre descubrimiento de verdaderos positivos y limitación de falsos positivos.

4.7. Escalado multidimensional

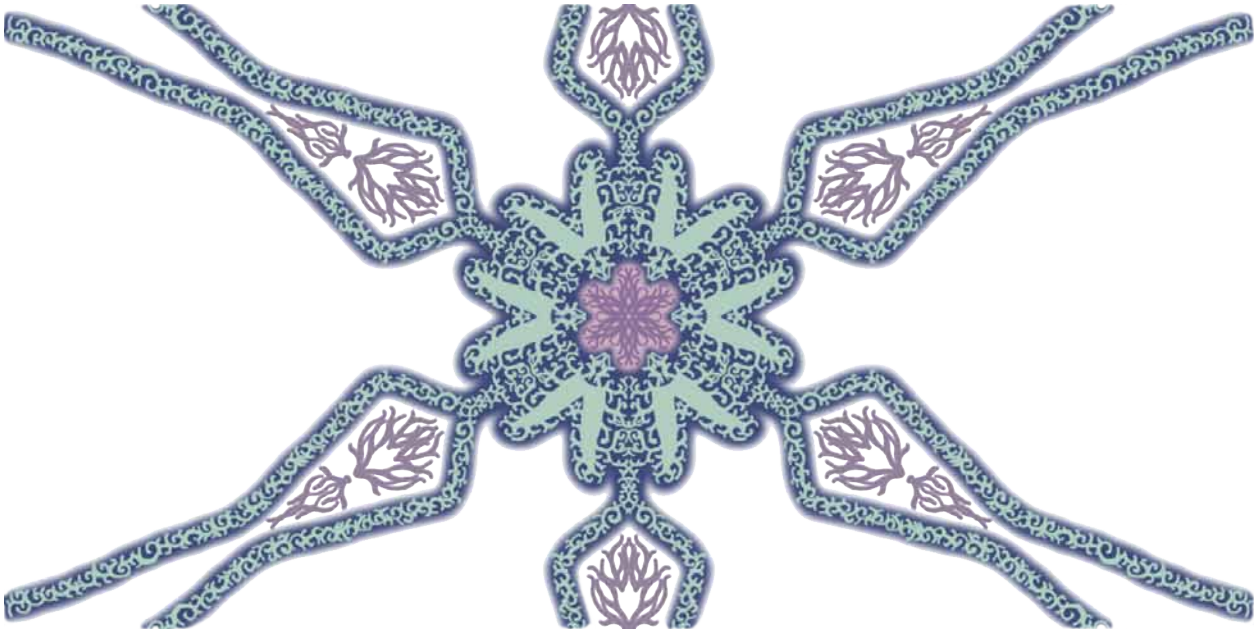
El escalado multidimensional de los perfiles se calculó con la función *cmdscale()* de la distribución básica de *R*.

There are two possible outcomes: if the result confirms the hypothesis, then you've made a measurement. If the result is contrary to the hypothesis, then you've made a discovery.

Enrico Fermi (1901 – 1954)

5

Resultados y Discusión



En este capítulo se mostrarán los resultados obtenidos al analizar las relaciones entre las anotaciones funcionales en la ontología GO y aquellas anotaciones fenotípicas en CMPO de un conjunto de genes silenciados mediante experimentos de RNAi.

5.1. Estudio de la densidad de anotaciones gen-fenotipo

La matriz que contiene las anotaciones de los genes a los fenotipos –extraídas de los experimentos detallados en la Tabla 4.1– está formada por 4198 genes y 36 fenotipos. Una representación visual de dicha matriz se encuentra en la Figura 5.1.

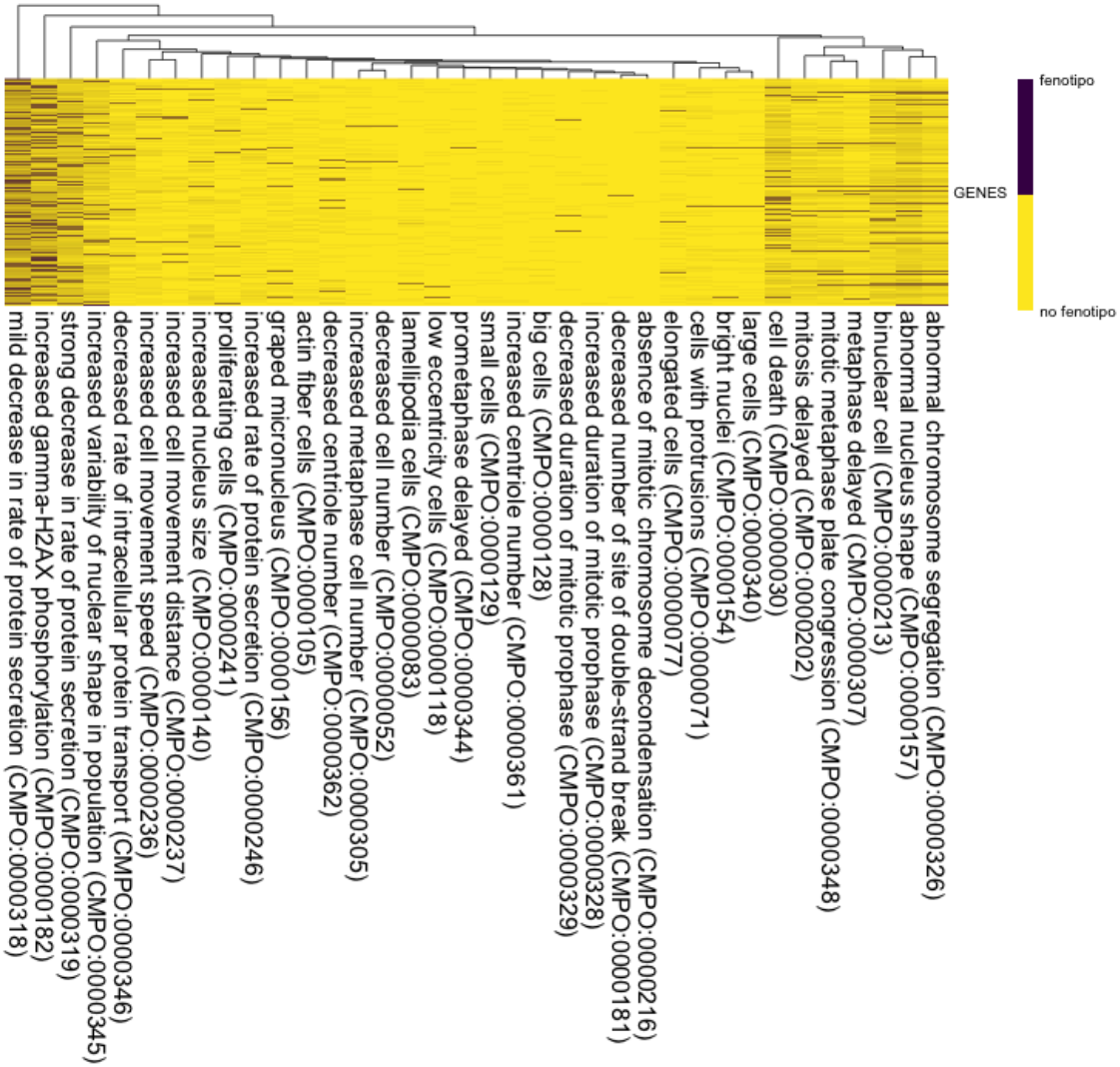


Figura 5.1: **Mapa de calor de la matriz gen-fenotipo.** Cada fila del mapa de calor representa un gen y cada columna un fenotipo. Las columnas están ordenadas según la correlación entre los fenotipos, como indica el dendrograma de la parte superior. Al ser la matriz binaria, sólo aparecen dos valores en la leyenda (fenotipo/no fenotipo).

Dadas las dimensiones de la matriz ($4198 * 36$), ésta podría albergar un total de 151128 posibles anotaciones. Sin embargo, sólo 7533 anotaciones gen-fenotipo aparecen en los experimentos (menos del 5 %), lo que indica una baja densidad de anotaciones en la matriz, un efecto también apreciable en la predominancia del color amarillo de la figura 5.1.

Analizando la distribución del número de anotaciones de fenotipos por gen (Figura 5.2), se observa un rango de entre 1 y 15, aunque la media es de 1.8 (más datos en la Tabla 4.3). De los 4198 genes, más de la mitad (2540) presentan un único fenotipo.

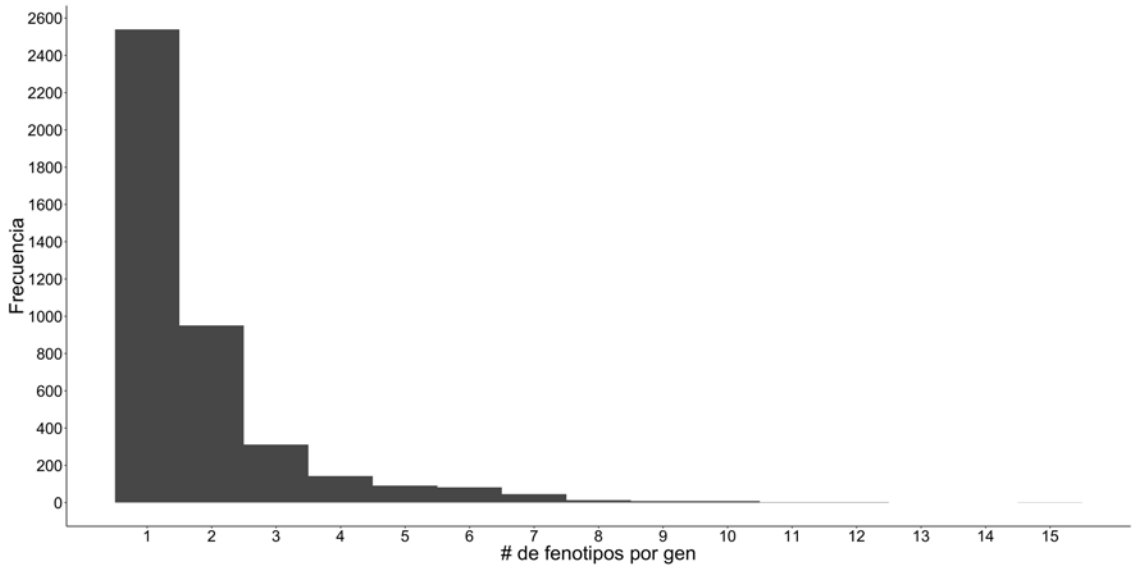


Figura 5.2: **Distribución del número de fenotipos por gen.** Representación de la frecuencia (eje Y) del número de fenotipos por gen (eje X). De los 4198 genes, más de la mitad (2540) están anotados en un solo fenotipo.

Este histograma nos muestra una matriz de perfiles fenotípicas muy poco densa donde alrededor de un 80 % de los genes están anotados a uno o dos fenotipos.

5.2. Estudio de la distribución de las anotaciones de genes en GO

La selección de genes de la matriz se ha efectuado a partir de los experimentos detallados en la Tabla 4.1, poniendo atención en que éstos cubran un rango amplio de procesos celulares. De esos genes, sólo se han seleccionado los 4198 genes que tienen anotación tanto en GO como en CMPO, es decir, para este estudio únicamente se han incluido los que tienen evidencia funcional y fenotípica (Figura 5.3).

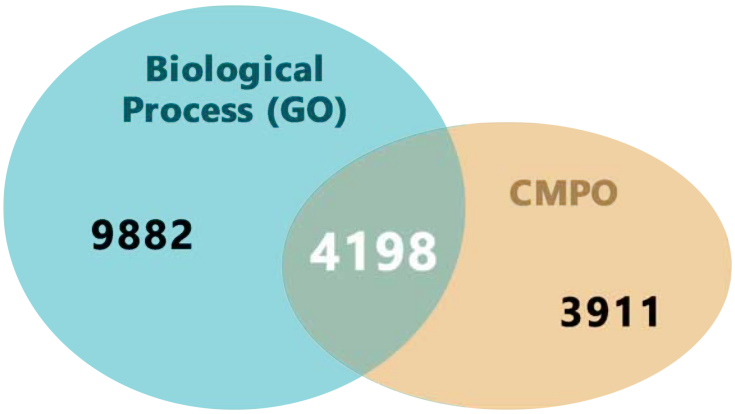


Figura 5.3: **Proceso de selección de genes.** En la rama *«cellular process»* de GO hay 14080 genes anotados y en CMPO 8109 genes. De todos ellos, 4198 comparten anotación tanto en GO como en CMPO.

Ahora bien, como se detalló en la sección 1.1.9, las anotaciones de los genes en GO pueden estar sesgadas, o lo que es lo mismo, los genes que presenta fenotipos podrían estar anotados en términos muy poco específicos de la ontología o bien estar presentes únicamente en términos muy específicos de una rama concreta. Esto podría deberse, por ejemplo, a genes que tradicionalmente son más estudiados por ser relevantes en el desarrollo de alguna enfermedad o en una función biológica.

En la rama *«cellular process»* de GO hay un total 14080 genes anotados, de los cuales 4198 también están anotados en CMPO (Figura 5.3). Por ello, para asegurar que los genes silenciados que muestran fenotipos no conforman un conjunto sesgado de anotaciones en GO, se compararon dos distribuciones del contenido informativo (IC, véase la sección 1.1.8).

Por un lado, se calculó la distribución del IC de todos los términos bajo la rama *«cellular process»* con los 14080 genes anotados (Figura 5.4, curva azul). Con esta distribución se manifiesta la forma en que todos los genes anotados en la rama *«cellular process»* de GO se distribuyen en cuanto a la especificidad de sus términos. Y por otro, se calculó la distribución del IC de los términos funcionales bajo *«cellular process»* en los que los 4198 genes con anotación fenotípica se anotan (Figura 5.4, curva violeta). Así se evalúa la especificidad de los términos en que está anotado el conjunto de genes que se está analizando.

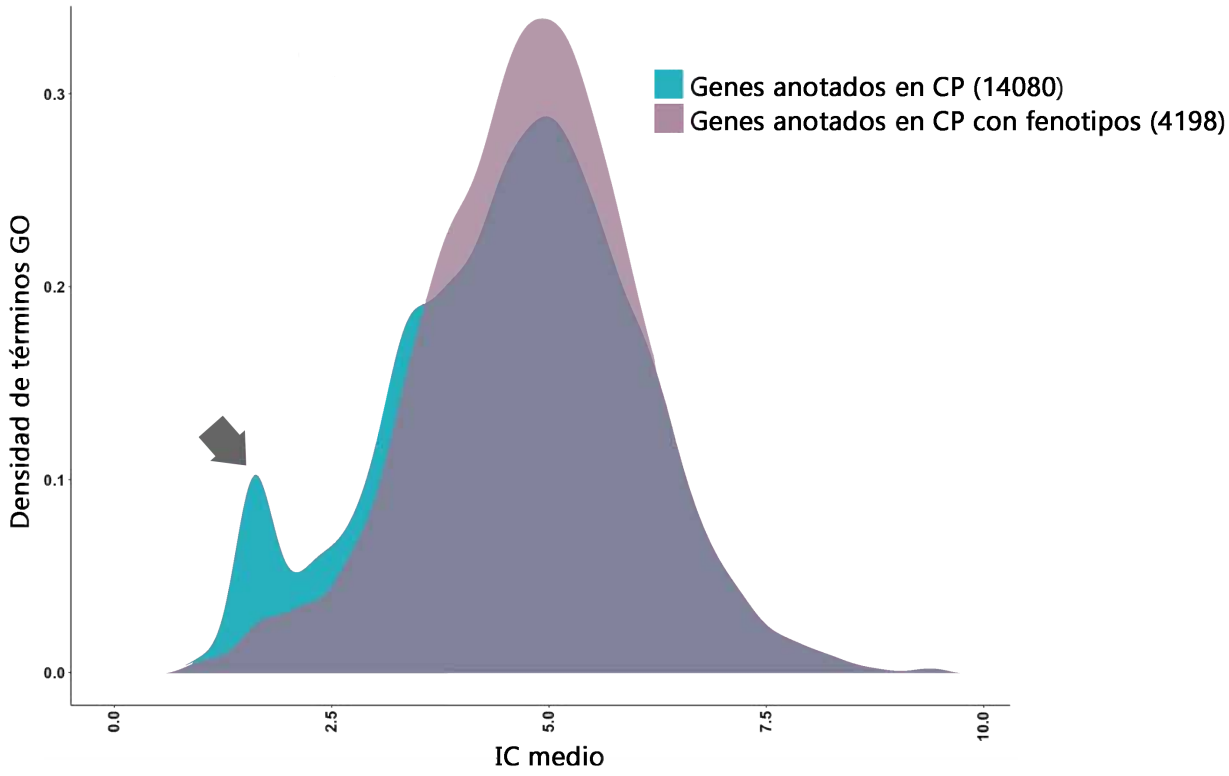


Figura 5.4: **Distribución del contenido informativo (IC) de los términos GO con genes anotados en fenotipos (morado) frente a todos los términos bajo *«cellular process»* (azul).** Las curvas representan la proporción de genes anotados en términos de *«cellular process»* (todas las anotaciones, en azul) frente al subconjunto de esos genes que presentan fenotipos (morado). Los valores incrementales del IC medio representan mayores valores de especificidad.

La comparación de ambas distribuciones permite conocer la especificidad de los términos en los que los genes están anotados, tanto en el conjunto de términos en estudio, como el total de genes anotados en la rama *«cellular process»* de GO.

Se observa en la Figura 5.4 que ambas distribuciones son muy similares, pero aquella que representa todos los genes anotados en *«cellular process»* (azul), presenta una densidad mayor para ciertos valores de IC bajos (IC $\leqslant$ 2.5, marcado con una flecha en la Figura 5.4). Una posible causa de este fenómeno es el problema que se mencionó en la sección 1.1.9 sobre aquellos genes de los que no se tiene un conocimiento claro de su función y que terminan siendo anotados en términos muy poco específicos de la ontología. En el conjunto de genes de este estudio aparece también ese sesgo, aunque de forma menos acentuada, como se observa en la flecha de la Figura 5.4.

De todo ello se concluye que los 4198 genes seleccionados están anotados en la ontología siguiendo una distribución similar a la que siguen las anotaciones de GO y además con un menor sesgo en cuanto a genes anotados en términos funcionalmente muy generales (estos últimos correspondientes a la flecha de la Figura 5.4).

5.3. Estudio de dependencia entre las métricas de similitud fenotípica

Normalmente, la similitud entre perfiles se mide aplicando métricas basadas en vectores (descritas en la sección 1.1.6) como la *correlación* (Laufer et al., 2013; Bakal et al., 2007) o el *coseno* (Loo et al., 2007; Wang et al., 2012). Al ser vectores binarios, también se aplican métricas de similitud especializadas en caracteres binarios. Por ejemplo, el algoritmo *PhenoBlast* recupera perfiles similares a una consulta dada, esto es, el número de coincidencias en la cadena binaria (Gunsalus et al., 2004). *PhenoBlast* tiene en cuenta que hay algunos fenotipos que podrían ser más informativos que otros, pero también considera la probabilidad de observar una combinación dada de fenotipos por azar. Combinando X e Y se obtiene la similitud de *Cohen's kappa*.

El hecho de que algunos fenotipos sean más informativos que otros puede formalizarse usando similitudes basadas en el contenido informativo, ya explicadas en las secciones de métricas vectoriales (1.1.6) y ontológicas (1.1.8). En este contexto, el contenido informativo se refiere a la especificidad de un fenotipo dentro de un contexto de anotaciones. El criterio habitual para determinar que un fenotipo es específico es que se observe con poca frecuencia. Por ejemplo, la muerte celular, al ser un fenotipo muy observado (muy frecuente), se considera menos específico que «*mitotic delay*», este último observado en un menor número de casos (fenotipo menos frecuente).

Entre las métricas vectoriales que consideran la especificidad de los fenotipos se encuentra *TF-IDF* (Figura 1.10), que se basa en la teoría de la información. En ella, se asigna un peso a cada uno de los fenotipos que indica su frecuencia de ocurrencia en los datos (Groth et al., 2008). En el mismo sentido, la reciente publicación de la ontología CMPO ha permitido aplicar también métricas de similitud semántica (detalladas en la sección 1.1.8).

Tanto las métricas que se aplican sobre vectores, como aquellas que consideran la especificidad de los fenotipos, se han aplicado sobre la matriz original de genes y fenotipos. Sin embargo, cuando se trabaja en un espacio con un alto número de dimensiones, a veces es beneficioso calcular las similitudes basadas en vectores en un espacio de dimensiones más reducidas (más detalles en la sección 1.1.11.1). Como en este caso los perfiles fenotípicos son vectores que cons-

tan de 36 componentes o fenotipos distintos, es conveniente aplicar un método de reducción a esas 36 dimensiones. Al tratarse de vectores binarios, el más adecuado es PCA logístico (como se explicó en las secciones 1.1.11.1 y 4.2.2). Ya en el nuevo espacio reducido, se pueden calcular de nuevo algunas –no todas, como se justificó en la sección 1.1.11.1– de las métricas de similitud.

Una forma de representar los perfiles fenotípicos proyectados en un espacio bidimensional es el escalado multidimensional (MDS) (véase la sección 4.7). En la Figura 5.5, se puede ver la distribución de los genes en un espacio bidimensional de acuerdo a la distancia proporcionada por Resnik en CMPO. No se observan comunidades claras pero sí ciertos grupos que comparten fenotipos.

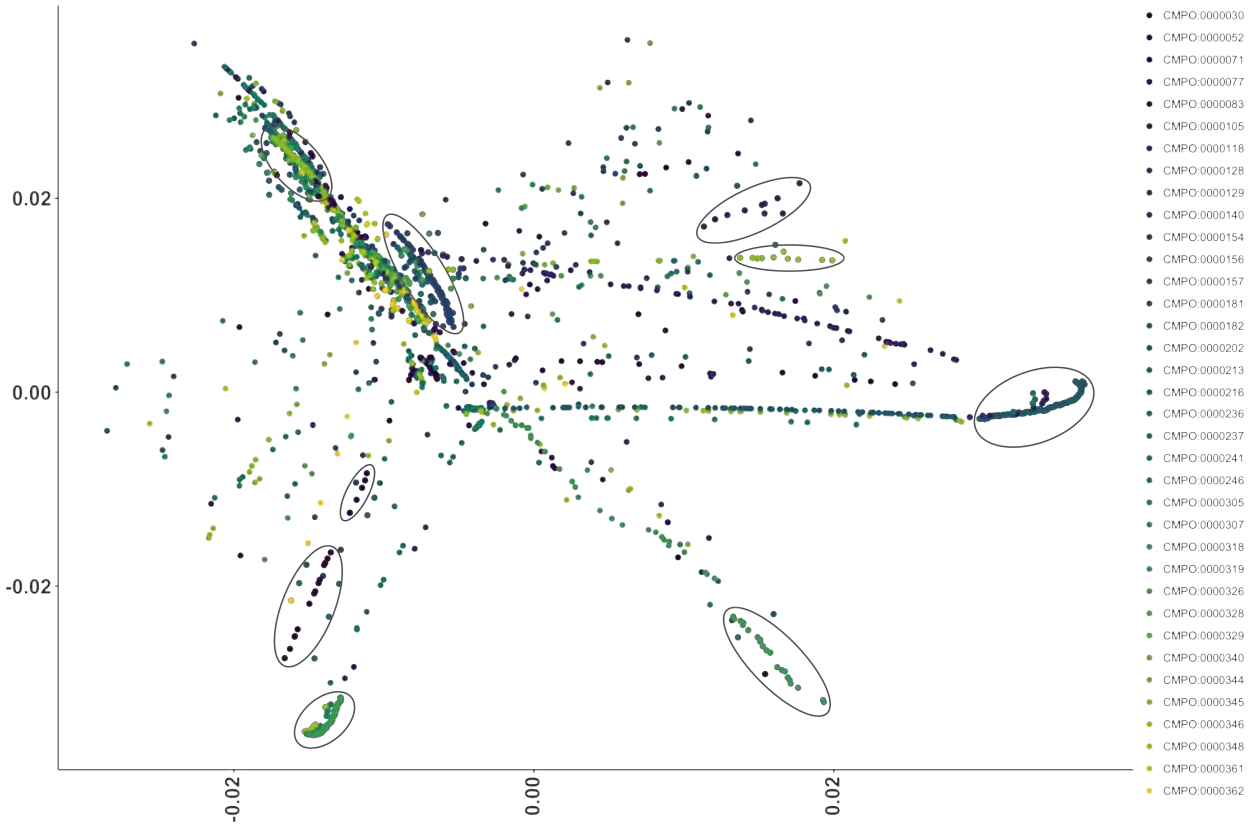


Figura 5.5: **Escalado multidimensional de genes para Resnik en CMPO.** Cada punto representa un gen y los colores indican fenotipos distintos, cuyos códigos CMPO aparecen en la leyenda. La distribución de los genes con perfiles multidimensionales en un espacio bidimensional pone de manifiesto algunos grupos (marcados con elipses) que comparten fenotipo.

Dados los distintos métodos con los que se puede medir la similitud fenotípica, la pregunta que surgió en este punto es si esas métricas serían ortogonales o co-dependientes entre sí. Para responderla, se calculó el coeficiente de correlación (PCC) entre las similitudes fenotípicas obtenidas –para todos los pares de genes– en las distintas métricas y se representó mediante un agrupamiento jerárquico (Figura 5.6).

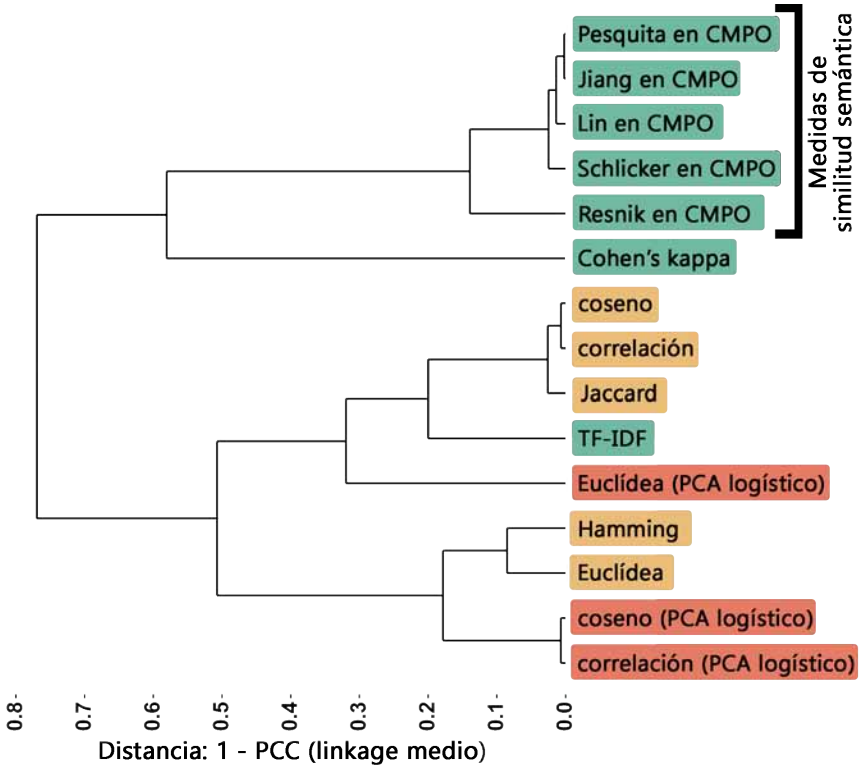


Figura 5.6: **Agrupamiento jerárquico de similitudes fenotípicas.** Dendrograma de métricas de similitud fenotípica en base a la distancia medida mediante la correlación de Pearson. En verde, las métricas que se basan en la teoría de la información; en amarillo las vectoriales; y en naranja aquellas aplicadas tras la reducción de dimensiones de la matriz.

El dendrograma resultante (Figura 5.6) muestra que las medidas de similitud se separan en dos grupos principalmente: por un lado aquellas que se basan en la similitud semántica (*Pesquita*, *Jiang*, *Lin*, *Schlicker* y *Resnik*) y por otro, las métricas basadas en los vectores fenotípicos (*coseno*, *Euclídea*, *correlación*, *Jaccard* y *Hamming*), con *Cohen's kappa* ocupando una posición intermedia, lo que confirma la idea de que estos grupos de métricas miden la similitud fenotípica de manera muy distinta.

5.4. Estudio comparativo de las métricas de similitud fenotípica

Hasta el momento se ha tenido en cuenta la información fenotípica, pero no la funcional. En esta sección se comparan e identifican las métricas que mejor correlacionan la similitud fenotípica y funcional entre genes.

Una forma habitual de comprobarlo consiste en considerar las interacciones entre proteínas como casos positivos de la relación funcional entre genes. Esto se traduce en que si dos proteínas interactúan, se puede considerar que los genes correspondientes participan en la misma función (Sharan et al., 2007). Es decir, para una métrica relevante, se espera que los genes similares fenotípicamente estén enriquecidos en interacciones entre proteínas.

Para comprobarlo se usaron tres aproximaciones. Por un lado, se comprobó la capacidad que tiene cada métrica de distinguir entre pares interactuantes y no interactuantes calculando el área bajo la curva ROC (AUC) (explicada en la sección 1.1.12). En una segunda aproximación, para cada similitud y gen se identificaron los vecinos más cercanos –los genes más similares fenotípicamente– y se comprobó si esos dos genes eran a su vez pares interactuantes. Finalmente, con el test estadístico de Mantel (secciones 1.1.12 y 4.4.2) se evaluó la correlación entre la matriz de interacciones entre proteínas y cada una de las matrices de similitud fenotípica.

5.4.1. Primera aproximación: Capacidad de cada métrica de distinguir entre pares interactuantes y no interactuantes

Las métricas de similitud fenotípica dan una idea de la cercanía entre perfiles y por ello, facilitan la agrupación de genes en función de sus fenotipos. Analizar la capacidad discriminatoria de dicha métrica significa estudiar la idoneidad a la hora de diferenciar genes que pertenecen a un conjunto de evaluación positivo y otro negativo en base a un criterio. En ese sentido, las redes de interacción de proteínas proporcionan el criterio para evaluar la métrica.

El significado del área bajo la curva ROC o AUC en este planteamiento es la probabilidad de que la similitud fenotípica dé un valor más alto a un par interactuante que a uno no interactuante. Es decir, los pares cercanos fenotípicamente se espera que sean pares interactuantes. Una similitud que no tenga capacidad de discriminar entre ambos conjuntos tendrá un AUC de 0.5 (correspondería a una distinción aleatoria de positivos y negativos); mientras que valores superiores indican una capacidad de discriminación que será más alta progresivamente.

En este contexto, es necesario definir un conjunto positivo y otro negativo. Como positivo, se usaron interacciones físicas proteína-proteína con alto grado de confianza. Para ello, se tomaron las interacciones físicas de cuatro fuentes distintas (véase la Sección 4.4.1) –`Intact` (Orchard et al., 2014), `MIPS` (Pagel et al., 2005), `DIP` (Salwinski et al., 2004) y `BIOGRID` (Stark et al., 2006)– que se combinaron manteniendo sólo aquellos pares identificados por al menos dos de esas fuentes o bien por ser interacciones curadas en `Reactome` (Milacic et al., 2012; Fabregat et al., 2016). Como conjunto negativo se usó un grupo de interacciones negativas curadas descritas en el `MIPS Negatome` (Blohm et al., 2014) y Trabuco et al (Trabuco et al., 2012).

Con el objetivo de verificar la idoneidad de estos dos conjuntos –control positivo y negativo de interacciones entre proteínas–, se calculó la similitud semántica de Resnik (SS) entre los pares de genes del conjunto de estudio anotados en GO. Posteriormente, se estimó la curva ROC basada en la métrica de SS y se calculó su correspondiente AUC. Con esta prueba, se evaluó la capacidad de Resnik para distinguir el grupo positivo –proteínas interaccionantes– y el negativo –proteínas no interaccionantes–, dando así una orientación del grado de correspondencia entre la similitud funcional y las interacciones físicas (Figura 5.7).

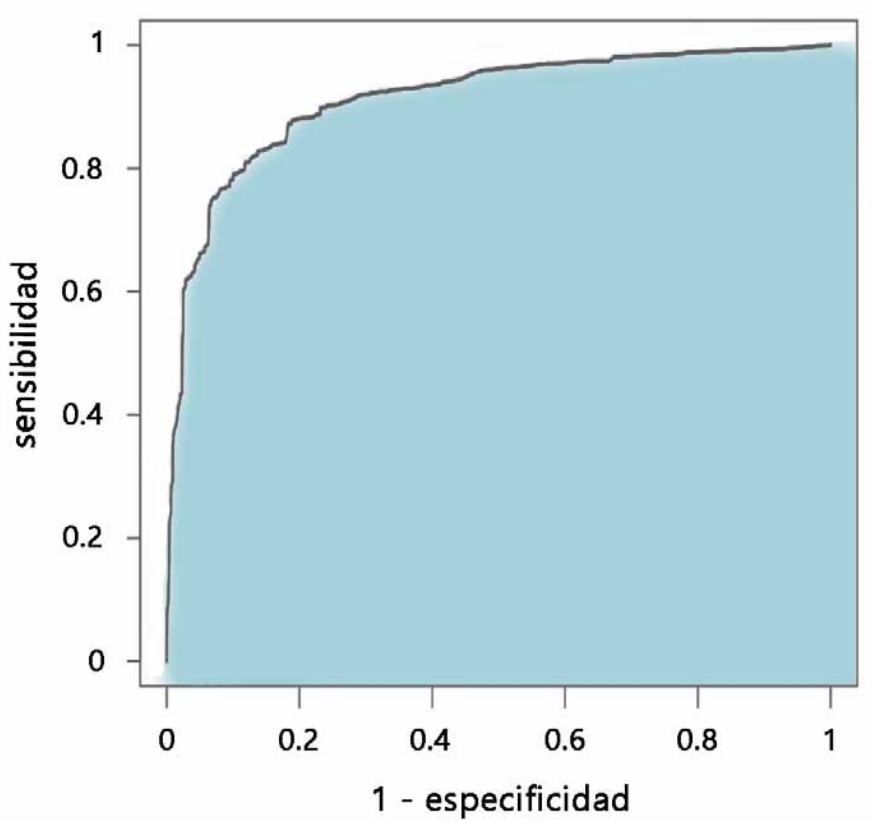


Figura 5.7: **Curva ROC para evaluar la similitud semántica en GO.** El área bajo la curva (AUC) es 0.913, lo que indica la buena capacidad de predicción de las interacciones entre proteínas a partir de la similitud semántica en GO.

El área bajo la curva es 0.913, lo que indica que la similitud semántica en GO tiene buena capacidad discriminatoria entre las interacciones de proteínas, o lo que es lo mismo, puede predecir con cierto grado de fiabilidad las interacciones físicas entre proteínas.

Una vez comprobada la estrecha relación funcional en GO de los genes codificantes de proteínas interaccionantes, se aplicó este mismo proceso de validación a las métricas de similitud fenotípica entre genes. De esta forma, se evalúa la capacidad discriminatoria entre positivos y negativos de cada una de las métricas que se comparan en esta sección. En la Tabla 5.1, se encuentran todas las similitudes fenotípicas ordenadas por su área bajo la curva (AUC).

De acuerdo con este criterio, las mejores similitudes son *Resnik*, *Schlicker* y *Lin*, siendo el resto de métricas menos adecuadas para reflejar el comportamiento funcional (Tabla 5.1).

Un factor que afecta a que las métricas sobre ontologías obtengan mejor resultado es el hecho de que en distintos experimentos existan diferencias en la descripción de los fenotipos. Por ejemplo, los términos *«metaphase delayed»* y *«mitosis delayed»* se tratan como fenómenos independientes en las métricas vectoriales. Aunque la relación entre ambos fenotipos parece evidente, cuando el cálculo se hace de forma vectorial, la similitud entre dos genes viene dada por los valores de esas dos características de forma separada. Sin embargo, cuando se usan las medidas de similitud semántica en la ontología, este problema desaparece, pues se elimina redundancia y se organizan jerárquicamente los términos.

SIMILITUD FENOTÍPICA	AUC
Resnik en CMPO	0.56
Schlicker en CMPO	0.56
Lin en CMPO	0.55
Cohen's kappa	0.54
Pesquita en CMPO	0.54
Jiang en CMPO	0.54
TF-IDF	0.53
Euclídea	0.53
correlación	0.52
Hamming	0.52
coseno	0.49
Jaccard	0.49
Euclídea (PCA logístico)	0.46
correlación (PCA logístico)	0.45
coseno (PCA logístico)	0.45

TABLA 5.1: **Medidas de similitud fenotípica ordenadas por el área bajo la curva ROC (AUC).** Un área de 0.5 indica que la métrica no tiene capacidad para discriminar entre pares interactuantes y no interactuantes. La capacidad de discriminar entre estos dos conjuntos aumenta conforme se incrementa el AUC.

La reducción de dimensiones tras el PCA no mejoró la respuesta de las métricas entre vectores (Tabla 5.1), lo que indica que las combinaciones lineales de fenotipos –causadas por la aplicación del PCA– no parecen plasmar de forma muy acertada los enlaces con la función de los genes. Por todo ello, cuando se usan las métricas de similitud semántica, los perfiles fenotípicos de genes interactuantes son en general más similares que para aquellos genes no interactuantes.

5.4.2. Segunda aproximación: Capacidad de cada métrica para detectar interacciones entre vecinos cercanos fenotípicamente

Se puede plantear la similitud fenotípica como los valores de los enlaces de una red de genes. De esta forma, los nodos (genes) más cercanos presentarán una similitud fenotípica alta. Si para cada gen, se identifica el vecino más cercano en dicha red, esto es, el gen más similar en cuanto a los fenotipos que ambos presentan, se puede comprobar si esos dos genes son, a su vez, pares interactuantes en una PPI (*iRef index* en este caso).

Para comparar las similitudes fenotípicas entre sí, se ordenaron por el número de interacciones que recuperaban (Tabla 5.2).

SIMILITUD FENOTÍPICA	INTERACCIONES EN PPI	P-VALOR
Cohen's kappa	27	0.0015
TF-IDF	25	0.0055
Resnik en CMPO	24	0.0102
correlación	22	0.0311
Hamming	21	0.0513
Euclídea	16	0.3433
Pesquita en CMPO	14	0.5494
coseno	13	0.6545
Jaccard	13	0.6545
Schlicker en CMPO	12	0.7512
Lin en CMPO	11	0.8332
Jiang en CMPO	11	0.8332

TABLA 5.2: **Medidas de similitud fenotípica ordenadas por el número de proteínas interactuantes recuperadas**. La primera columna representa el número total de vecinos más cercanos que además son pares interactuantes, mientras que la segunda indica el p-valor (calculado usando la distribución hipergeométrica) de que el número de observaciones interactuantes se deba al azar.

Con este método, *Cohen's kappa*, *TF-IDF* y *Resnik* son las que obtienen mejor rendimiento. Las métricas que calculan la probabilidad de ocurrencia de un fenotipo –como es el caso de TF-IDF y *Cohen's kappa*– aunque se basan en el cálculo entre vectores fenotípicos, son más útiles que el resto de medidas de similitud vectorial. Esto podría atribuirse a la relación entre la frecuencia de un fenotipo y su especificidad. Es decir, los fenotipos más específicos aparecerán menos en los datos y viceversa. Por esta razón, guardan una cierta relación con las métricas de similitud semántica en ontologías. Otras similitudes semánticas, junto con las vectoriales, no obtuvieron mejores resultados que una selección aleatoria de interacciones entre proteínas.

5.4.3. Tercera aproximación: Correlación entre las matriz de interacciones y las matrices de similitud fenotípica

Con el objetivo de calcular la correlación entre las distintas similitudes fenotípicas y las interacciones físicas entre proteínas, se aplicó el test estadístico de Mantel. Para ello, las interacciones entre proteínas procedentes de la base de datos *iRef index* se expresaron en formato matricial, donde la interacción se representó con el valor 1 y la ausencia de interacción con el valor 0. De forma análoga, se construyó una matriz de relaciones fenotípicas entre genes con los valores de similitud para cada métrica.

Posteriormente, aplicando el test estadístico de Mantel, se midió la correlación entre cada matriz de similitud fenotípica y la matriz de interacciones entre proteínas (Tabla 5.3).

SIMILITUD FENOTÍPICA	CORRELACIÓN	P-VALOR
Resnik en CMPO	$4{,}81e-03$	0.005
TF-IDF	$1{,}64e-03$	0.103
Hamming	$2{,}04e-04$	0.498
Euclídea	$1{,}59e-04$	0.522
correlación	$-9{,}39e-05$	0.533
coseno	$-5{,}58e-04$	0.629
Jaccard	$-5{,}77e-04$	0.643

TABLA 5.3: **Medidas de similitud fenotípica ordenadas por la correlación con la matriz de interacciones.** La correlación mide el grado de correspondencia entre la métrica en cuestión y la matriz de interacciones de *iRef index*.

En los resultados se observan valores muy bajos de correlación. Esto se debe a que la matriz de interacciones es extremadamente dispersa. No obstante, la métrica con mejor resultado es la similitud semántica de *Resnik*.

5.4.4. Selección de la métrica más adecuada atendiendo a las distintas aproximaciones

La primera aproximación de las tres aplicadas presenta un resultado débil, el número de interacciones totales es bajo, al igual que manifiesta el bajo valor de correlación entre las matrices de similitud fenotípica y la de interacciones entre proteínas. También en la aproximación de la curva ROC se obtuvieron AUCs muy cercanas a la aleatoriedad (0.5). No obstante, el único objetivo de estas tres aproximaciones es ordenar las métricas según su cercanía con la similitud funcional.

En general, parece que *Resnik* es la más consistente cuando se trata de asociar fenotipos similares con proteínas interactuantes, pues aparece en los tres casos en las primeras posiciones.

El resto de similitudes semánticas pueden haberse visto influenciadas negativamente por la dispersión de la ontología CMPO (Tabla 5.4), ya que tienen más en cuenta la topología que *Resnik*. Como ya se indicó en la sección 1.1.8, *Lin* (Lin, 1998) y *Jiang* (Jiang and Conrath, 1997) son particularmente sensibles a las variaciones en la ontología, no así *Resnik* (Resnik, 1995). Esto significa que para un par de términos, *Resnik* calcula el IC del ancestro común más informativo (MICA), sin considerar la distancia a la que dicho ancestro se encuentra del par de términos. Sin embargo, *Lin* y *Jiang* sí tienen en cuenta esa distancia, por lo que estos dos métodos son más sensibles a las variaciones de la ontología. Al ser CMPO una ontología pequeña y con pocos niveles (Tabla 5.4), *Resnik* parece ser la medida comparativamente más robusta.

Descripción	**Valor**
Número de términos	361
Número de relaciones entre términos	391
Número de anotaciones	13866
Número de genes distintos anotados	8109
Densidad del grafo (número de relaciones/posibles relaciones)	0.00422
Número de islas	1
Diámetro de la ontología	6

TABLA 5.4: **Estadísticas sobre la ontología CMPO**. La ontología CMPO tiene anotados fenotípicamente genes procedentes de humano y ratón.

5.5. Predicción de la relación funcional en GO a partir de la fenotípica en CMPO

Una vez decidida la métrica más adecuada para medir la similitud entre perfiles fenotípicos, se exploró cómo las funciones de los genes se relacionan con los experimentos de una forma más directa. Si los fenotipos predicen las funciones biológicas, se espera que los pares de genes con fenotipos similares desempeñen funciones similares.

Como las funciones de los genes han sido estandarizadas usando GO, la similitud funcional entre genes se calculó mediante la similitud semántica de *Resnik* entre términos GO usando el máximo para comparar términos en los que los genes están anotados (sección 1.1.8). Este método parece ser el más adecuado para calcular la similitud en GO (Guzzi et al., 2012).

Para evaluar la relación entre la similitud fenotípica y la similitud funcional de los pares genes, se representó gráficamente *Resnik* en GO frente a *Resnik* en CMPO para todos los pares de genes obtenidos en los experimentos de silenciamiento génico usando RNAi, excluyendo aquellos genes sin anotación funcional en GO (Figura 5.8).

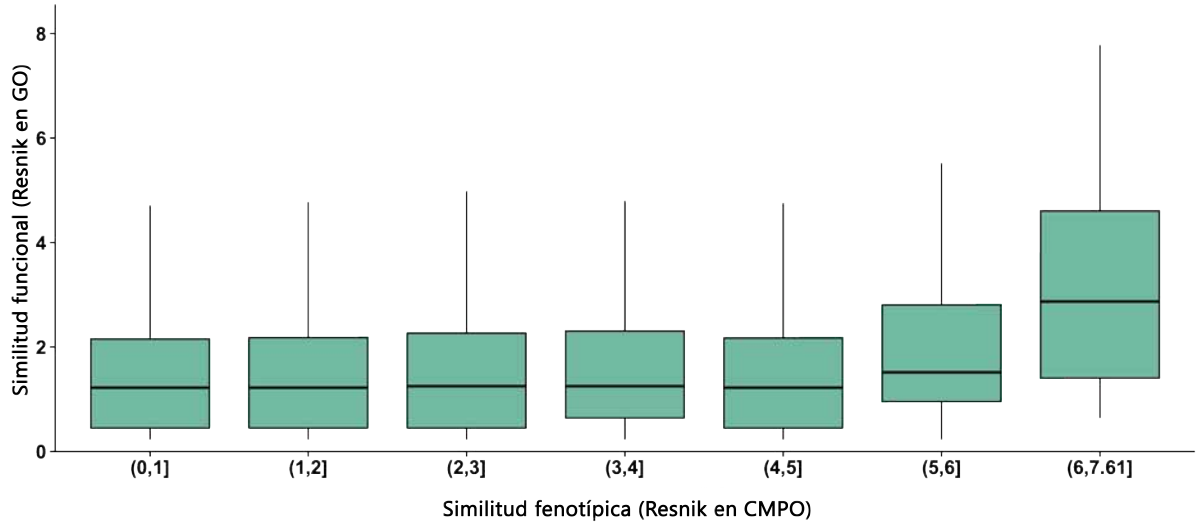


Figura 5.8: **Distribucion de la similitud fenotípica en CMPO frente a la similitud funcional en GO para todos los pares de genes.** Cada caja representa los márgenes del cuartil superior e inferior, con la mediana representada por la línea negra dentro de cada caja.

La distribución de los valores de similitud funcional obtenida es la misma para todos los niveles de similitud fenotípica, excepto para valores altos correspondientes al rango (6, 7.61], donde aparece una leve tendencia hacia valores también altos de similitud funcional. Esto significa que pares de genes muy similares fenotípicamente, lo son también funcionalmente. Sin embargo, esta tendencia es leve y se esperaría una correlación más clara entre ambas variables.

Esta ausencia de correlación es consistente con los resultados mostrados previamente: hay una alta correlación entre las redes de interacción de proteínas y GO, pero una baja correlación entre las redes de interacción de proteínas y CMPO. Por tanto, entre GO y CMPO no se espera un alto valor de correlación.

Por citar un caso concreto en que genes involucrados en un mismo proceso den lugar a fenotipos muy distintos, esto podría suceder con los genes que regulan positivamente o negativamente a otro. En GO pertenecerían al mismo proceso celular, pero fenotípicamente el resultado ha de ser muy distinto. Por ejemplo, los términos funcionales *«positive regulation of stem cell differentiation»* y *«negative regulation of stem cell differentiation»* son dos términos hijos en GO del término *«regulation of stem cell differentiation»*. Si un gen está anotado en el término de regulación positiva y otro en el término de regulación negativa, el término común a ellos es *«regulation of stem cell differentiation»* que al ser muy específico, hará que la similitud semántica entre los dos genes mencionados sea muy alta. Sin embargo, dado que sus efectos son contrarios, los vectores fenotípicos serán muy distintos entre ellos. Para paliar este sesgo, algunas aproximaciones recientes apuntan a la incorporación de información adicional a la estructura del grafo con una relación entre términos del tipo *«opposite-of»* (Sebastian et al., 2017).

5.5.1. Del fenotipo a la función: Análisis de la tendencia observada

La leve tendencia que aparece para altos valores de similitud funcional podría deberse al azar. Para comprobar si esto es así, se realizaron 1000 asignaciones aleatorias de los valores con alta similitud en CMPO, esto es, aquellos con alta similitud fenotípica procedentes del rango (6,7.61]. La media de la distribución original observada en la última caja de la Figura 5.8 está en 2.98 (Figura 5.9, línea azul). Cuando se aleatorizan los valores asignados a cada par de genes se obtiene la distribución de la Figura 5.9. La media tras la aleatorización se sitúa en 1.53 (línea naranja). Esto indica que la tendencia observada en la Figura 5.8 no parece ser efecto del azar.

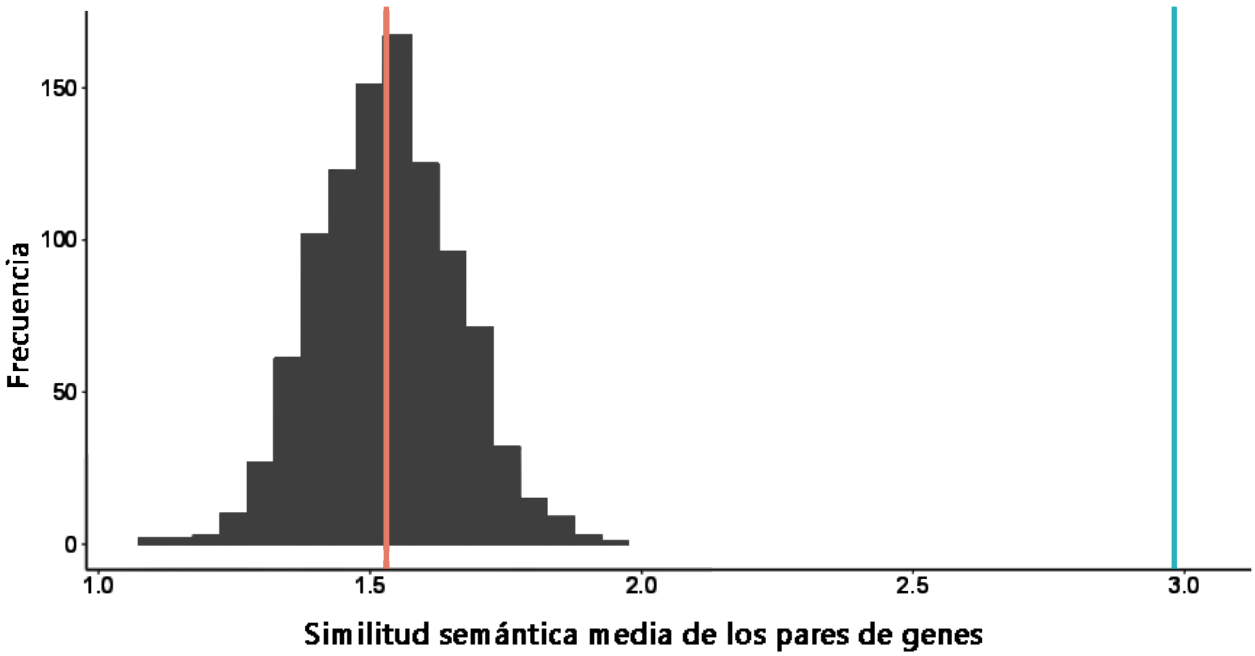


Figura 5.9: **Distibución aleatoria de similitud semántica media entre genes para aquellos pares con similitud semántica alta (>6).** Aleatorizando los valores de similitud fenotípica entre pares se obtuvo una media de 1.53 (naranja), con la distribución que se muestra en gris oscuro. Sin embargo, la media del conjunto original es de 2.98 (azul).

Sabiendo que la tendencia obtenida no se debe al azar, cabe preguntarse cuántos genes hay en esta distribución. Para el rango (6,7.61] se obtienen 20 genes distintos, listados en la Tabla 5.5.

GEN	NOMBRE DE LA PROTEÍNA	FUNCIÓN
OR4D1	Olfactory receptor 4D1	Receptor olfatorio.
AKAP5	A-kinase anchor protein 5	Anclaje de la proteína PKA al citoesqueleto y/o proteínas asociadas a organelos.
MC5R	Melanocortin receptor 5	Receptor de MSH y ACTH. La actividad de este receptor está mediada por proteínas G que activan la adenilato ciclasa.
RNF8	E3 ubiquitin-protein ligase RNF8	Señalización del daño del DNA.
RNF168	E3 ubiquitin-protein ligase RNF168	Acumulación de proteínas de reparación del DNA.
NUP153	Nuclear pore complex protein Nup153	Componente del complejo de poros nucleares, necesario para el tráfico a través de la membrana nuclear. Funciona como andamiaje para el transporte nucleocitoplasmático normal de proteínas y mRNAs.
MDC1	Mediator of DNA damage checkpoint protein 1	Detención del ciclo celular mediado por el punto de control en respuesta al daño del DNA.
DNMT3B	DNA (cytosine-5)-methyltransferase 3B	Metilación de novo de todo el genoma, esencial para el establecimiento de patrones de metilación del ADN durante el desarrollo.
TOP2A	DNA topoisomerase 2-alpha	Control de los estados topológicos del DNA por rotura transitoria y posterior unión de las cadenas de DNA. Esencial durante la mitosis y meiosis.
SMC1A	Structural maintenance of chromosomes protein 1A	Cohesión cromosómica durante el ciclo celular y en la reparación del DNA.
SIAH1	E3 ubiquitin-protein ligase SIAH1	Mediación de la ubiquitinación y posterior degradación en el proteosoma de las proteínas diana.
KAT6A	Histone acetyltransferase KAT6A	Acetilación de residuos de lisina en las histonas H3 y H4.
NAA10	N-alpha-acetyltransferase 10	Crecimiento y desarrollo vascular, hematopoyético y neuronal.
POLR3C	DNA-directed RNA polymerase III subunit RPC3	Catalizador de la transcripción del ADN en ARN.
NEK6	Serine/threonine-protein kinase Nek6	Progresión del ciclo celular mitótico: segregación cromosómica en transición metafase-anafase, formación de huso mitótico y citocinesis.
RUVBL2	RuvB-like 2	Regulación transcripcional, replicación del ADN y probablemente reparación del ADN.
INTS1	Integrator complex subunit 1	Transcripción de pequeños RNAs nucleares.
H1FNT	Testis-specific H1 histone	Espermatogénesis normal y fertilidad masculina.
GINS1	DNA replication complex GINS protein PSF1	Inicio de la replicación del ADN y progresión de las horquillas de replicación del ADN.
CDK1	Cyclin-dependent kinase 1	Control del ciclo celular eucariota modulando el ciclo del centrosoma, así como el inicio mitótico.

TABLA 5.5: **Grupo de 20 genes con alta similitud fenotípica y funcional.** La información de estos genes ha sido extraída de UniProtKB (`www.uniprot.org`).

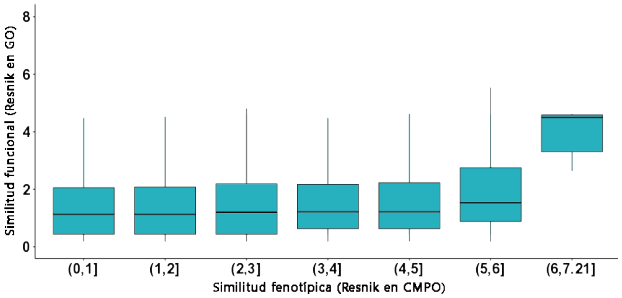
Estos 20 genes listados parecen estar involucrados en funciones muy dispares sin un enriquecimiento claro en ningún término GO específico.

Por tanto, teniendo en cuenta que el conjunto global consta de 4198, la proporción de genes con fenotipos muy específicos vinculados con funciones específicas es mínima. Se concluye de aquí que la mayoría de los fenotipos no parece ser buen indicador de la función biológica.

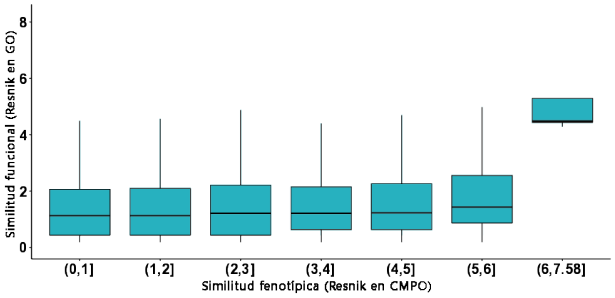
5.5.2. Del fenotipo a la función: Estudio de la robustez de la tendencia observada

Para comprobar si la dispersión de la matriz pudiera estar afectando a la señal obtenida, se eliminaron gradualmente anotaciones de la matriz gen-fenotipo. De esta forma, si la señal –aunque débil– se debiera al efecto de unos pocos genes, al ir eliminando anotaciones se observaría que la señal desaparece.

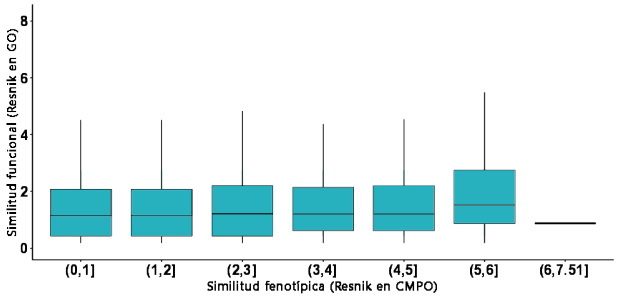
Se aprecia, sin embargo, que el crecimiento para los valores altos de similitud fenotípica es robusto, ya que se mantiene al eliminar el 5%, 10%, 20% y 30% de las anotaciones (panel de Figuras 5.10).



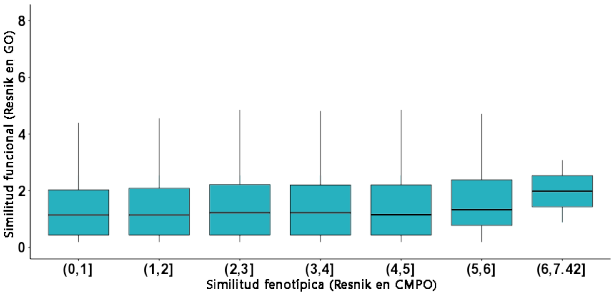
(a) 5% de las anotaciones eliminadas de la matriz original.



(b) 10% de las anotaciones eliminadas de la matriz original.



(c) 20% de las anotaciones eliminadas de la matriz original.



(d) 30% de las anotaciones eliminadas de la matriz original.

Figura 5.10: **Distribución de la similitud fenotípica en CMPO frente a la similitud funcional en GO filtrando anotaciones para todos los pares de genes.** Se mantiene la tendencia creciente en las últimas distribuciones para similitudes fenotípicas altas.

5.5.3. Del fenotipo a la función: Evaluación de la influencia de las anotaciones con evidencia electrónica

Como se comentó en la sección 1.1.7.1, las evidencias con las que pueden anotarse los genes en la ontología son muy distintas. Un caso particular son las anotaciones inferidas electrónicamente (IEA), muchas de ellas usadas para identificar anotaciones efectuadas mediante similitud de secuencia (Pesquita et al., 2008). Aunque esto podría producir circularidad en el estudio, algunos autores (véase la Sección 1.1.9) consideran que en ciertos casos estas anotaciones pueden introducir ruido, pero en general, su efecto suele ser positivo o nulo (Guzzi et al., 2012).

Para este caso concreto, la idea es evaluar si al incluir dichas anotaciones de GO la tendencia mejora, hecho que podría darse ya que al incluir IEA algunas relaciones podrían verse reforzadas por genes ortólogos. Sin embargo, los cambios apenas son perceptibles, como se observa al comparar las Figuras 5.8 y 5.11. Por tanto, las evidencias IEA no afectan a los resultados obtenidos de un modo destacable.

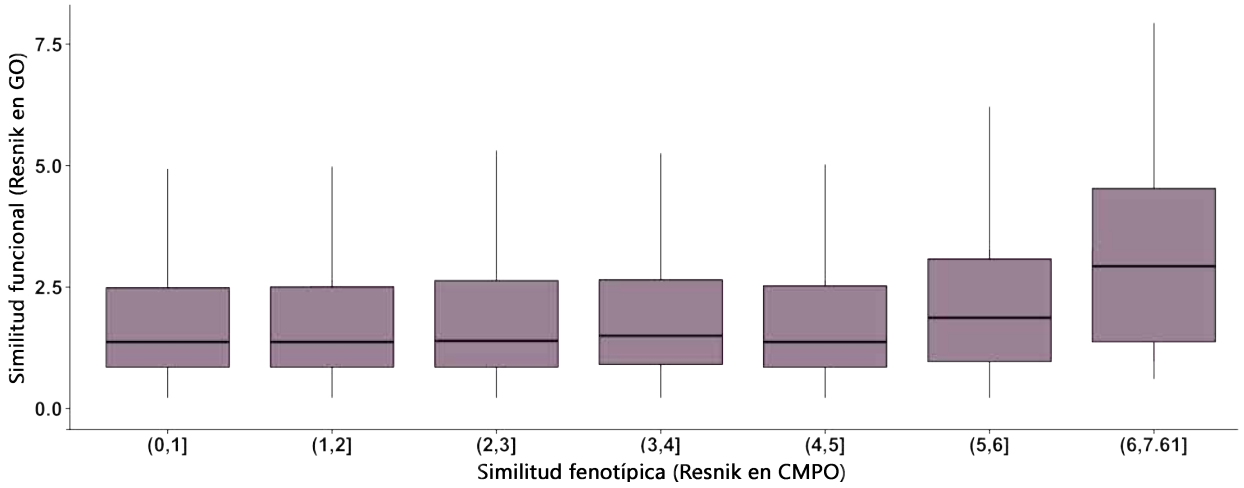


Figura 5.11: **Similitud fenotípica frente a la similitud funcional en GO incluyendo evidencias IEA.** Se sigue apreciando la tendencia creciente para el intervalo (6,7.61], pero no hay cambios destacables con respecto a lo obtenido en la Figura 5.8.

5.5.4. Del fenotipo a la función: Estudio de la influencia de los pares que no interactúan físicamente

Dado que las interacciones entre proteínas se han tenido en cuenta como casos positivos de relación funcional entre los genes, se podría pensar que si se mantienen solo aquellas similitudes fenotípicas que tienen verificada una interacción física, la tendencia puede hacerse más pronunciada. Para ello, se eliminaron los pares que no interactúan físicamente en la red de interacción de proteínas.

En la Figura 5.12 puede verse que esto no sucede, aunque en general sí aumentan levemente los niveles de similitud semántica en GO. Se sigue apreciando la tendencia creciente y aumentan los valores de similitud semántica en GO, pero la señal sigue siendo insuficiente para establecer una correlación entre ambas métricas.

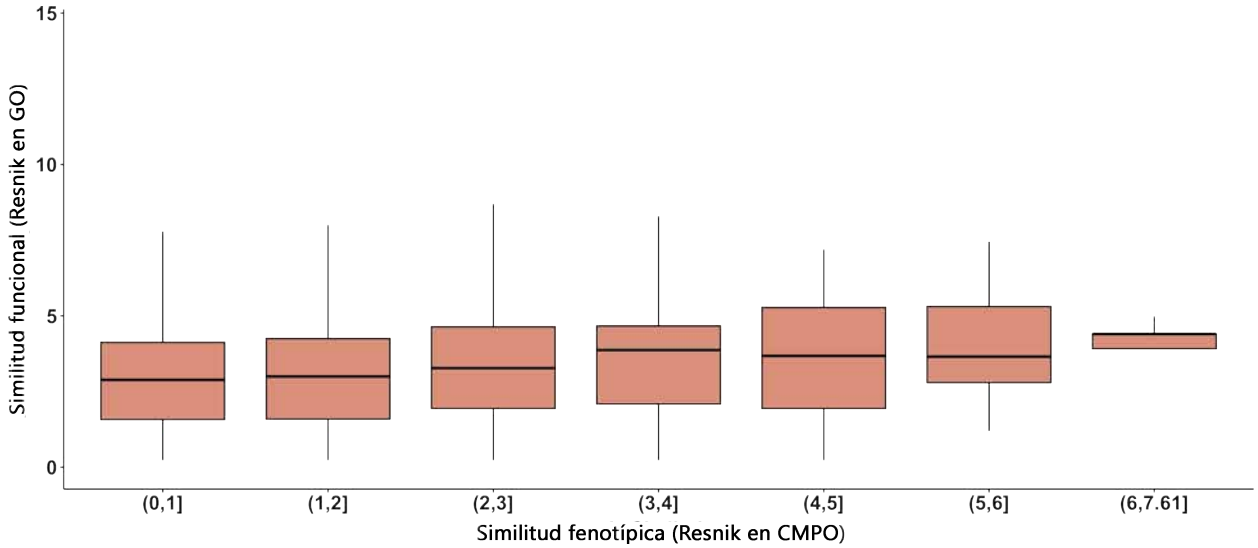
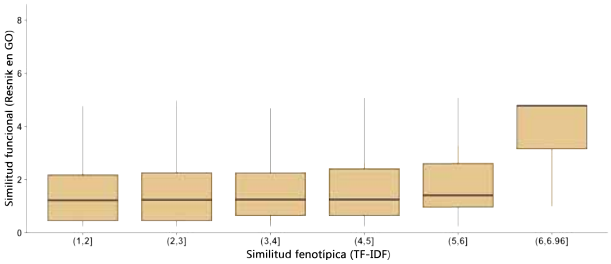


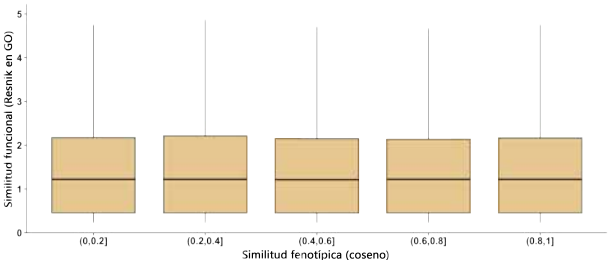
Figura 5.12: **Distribución de la similitud fenotípica en CMPO frente a la similitud funcional en GO filtrando los pares interactuantes en *iRef index*.** La tendencia sigue siendo débil.

5.5.5. Del fenotipo a la función: Resultados para las métricas de similitud vectorial

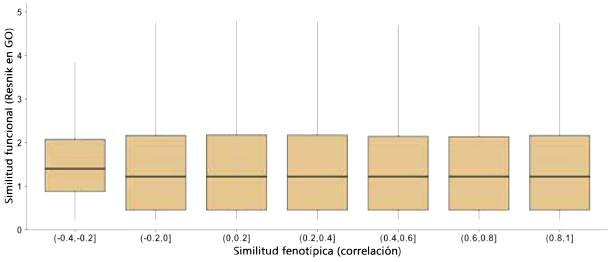
Hasta este punto, las distribuciones mostradas han usado, por un lado, la similitud semántica en GO –por ser la más adecuada para medir la similitud funcional entre genes–; y por otro, la similitud fenotípica aprovechando la estructura semántica que proporciona la ontología CMPO. En la sección 1.1.6 se vieron otras métricas que calculan la similitud entre perfiles fenotípicos. La comparativa entre la similitud funcional y fenotípica para estas otras métricas se encuentra en la Figura 5.13.



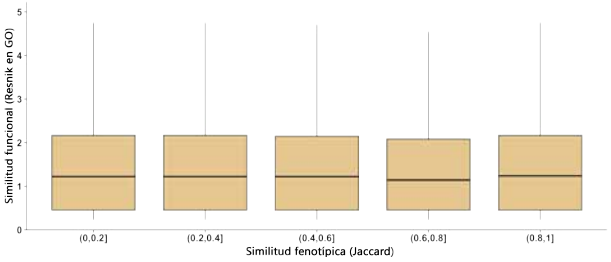
(a) Similitud fenotípica TF-IDF frente a la similitud funcional en GO.



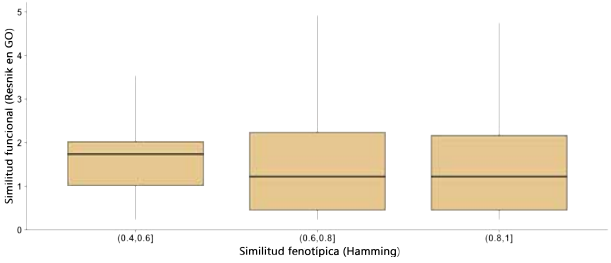
(b) Similitud fenotípica usando el coseno frente a la similitud funcional en GO.



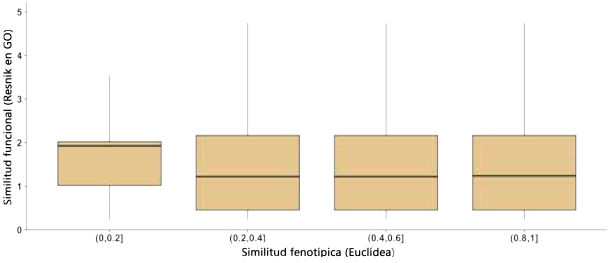
(c) Similitud fenotípica mediante correlación frente a la similitud funcional en GO.



(d) Similitud fenotípica Jaccard frente a la similitud funcional en GO.



(e) Similitud fenotípica Hamming frente a la similitud funcional en GO.



(f) Similitud fenotípica Euclídea frente a la similitud funcional en GO.

Figura 5.13: **Similitud fenotípica para distintas métricas vectoriales frente a la similitud funcional en GO.** La única métrica que muestra una tendencia mínima es TF-IDF.

La única métrica que muestra una tendencia débil es TF-IDF, el resto no presenta señal alguna. De hecho, el comportamiento de TF-IDF es muy similar al que muestra la similitud fenotípica de Resnik en CMPO. El motivo es que ambas funcionan de forma similar: Resnik en CMPO selecciona el término común más específico y TF-IDF selecciona el fenotipo común con mayor IDF, o lo que es lo mismo, más específico.

5.6. Predicción de la anotación fenotípica en CMPO a partir de la funcional en GO

De la sección anterior se deduce que sólo valores altos de similitud fenotípica se correlacionan con valores altos de similitud funcional. Esta tendencia es, además de consistente, robusta. No obstante, la señal es muy débil al implicar a pocos genes con valores de similitud semántica muy altas en GO y CMPO.

Una posible explicación es que varias funciones compartan el mismo fenotipo. Es decir, se obtiene un único fenotipo cuando se silencian genes distintos funcionalmente. A modo de ejemplo, esta afirmación tiene sentido cuando se silencia un gen del ciclo celular y otro involucrado en la secreción de proteínas. Aunque ambos procesos son distintos, pueden dar lugar al mismo fenotipo como puede ser la muerte celular.

Si este fuera el caso, entonces podríamos predecir que funciones similares pueden dar lugar a fenotipos similares. Esperaríamos entonces que dos genes involucrados en el mismo proceso celular presenten fenotipos similares. Como puede verse en la Figura 5.14, esto no ocurre: los genes con alta similitud funcional no suelen tener alta similitud fenotípica.

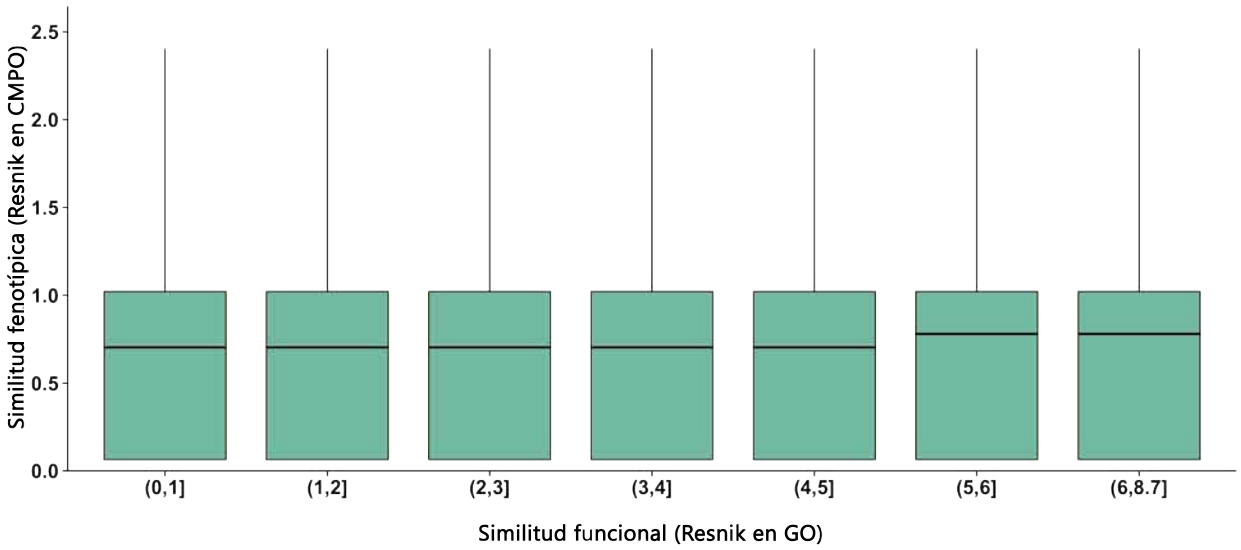
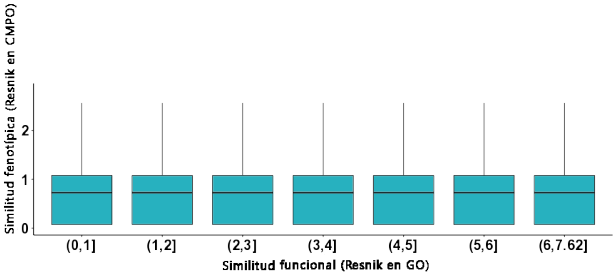


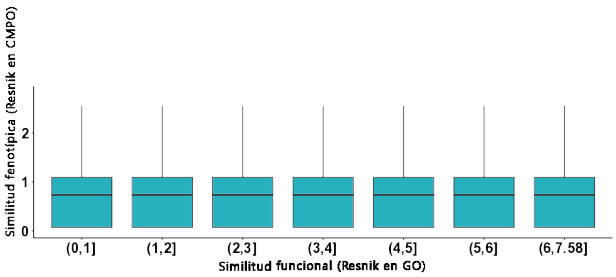
Figura 5.14: **Distribución de la similitud funcional en GO frente a la similitud fenotípica en CMPO para todos los pares de genes.** Para todos los rangos de similitud funcional en GO se obtienen distribuciones análogas de similitud fenotípica usando Resnik en CMPO.

5.6.1. De la función al fenotipo: Estudio de la robustez de las anotaciones

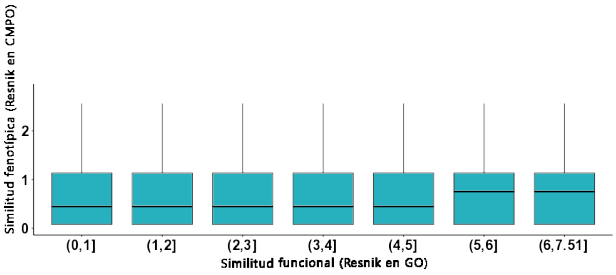
Al igual que en la sección 5.5.2, exploramos si al eliminar anotaciones de la matriz gen-fenotipo se obtienen resultados similares. De nuevo, los resultados se mantuvieron constantes (Figura 5.15).



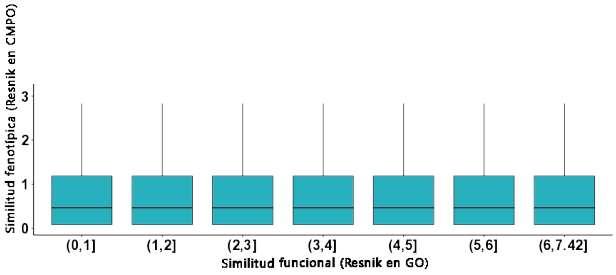
(a) 5% de las anotaciones eliminadas de la matriz original.



(b) 10% de las anotaciones eliminadas de la matriz original.



(c) 20% de las anotaciones eliminadas de la matriz original.



(d) 30% de las anotaciones eliminadas de la matriz original.

Figura 5.15: **Distribución de la similitud funcional en GO frente a la similitud fenotípica en CMPO eliminando parte de las anotaciones.** La tendencia entre la similitud funcional y la fenotípica sigue siendo inexistente.

5.6.2. De la función al fenotipo: Evaluación de la influencia de las anotaciones con evidencia electrónica

Podría darse la situación de que los genes anotados por inferencia electrónica (IEA) estuvieran introduciendo ruido en los datos. Sin embargo, en la ontología GO no se observó ningún cambio en las distribuciones de similitud fenotípica para distintos rangos de similitud funcional.

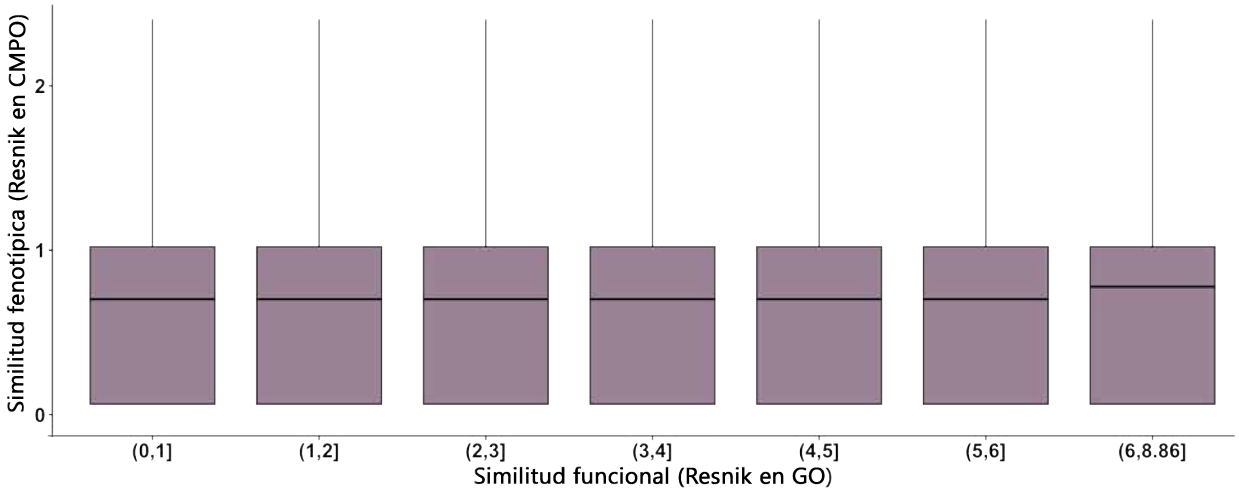


Figura 5.16: **Distribución de la similitud funcional en GO frente a la similitud fenotípica en CMPO incluyendo evidencias IEA.** No hay indicios de correlación alguna entre ambas variables.

5.6.3. De la función al fenotipo: Estudio de la influencia de los pares que no interactúan físicamente

De forma análoga al experimento de la sección 5.5.4, aquí se analizó el efecto de mantener únicamente aquellos pares que interactúan en *iRef index*, para comprobar si aparecía algún tipo de tendencia al filtrar los pares de genes. Los resultados son muy similares (véase la Figura 5.17), aunque de nuevo, el nivel general de similitud semántica aumenta, lo que indica que mantener solo los pares interactuantes no supone ninguna mejora.

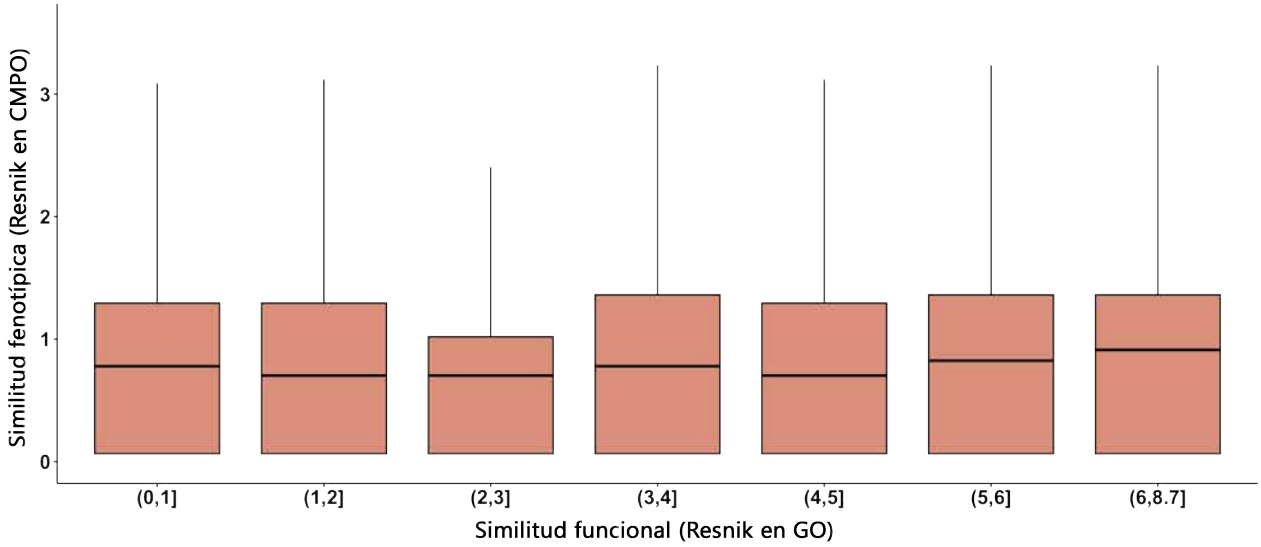
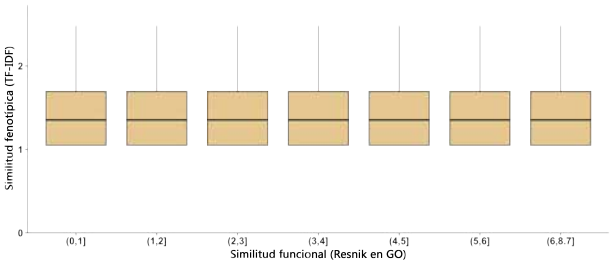


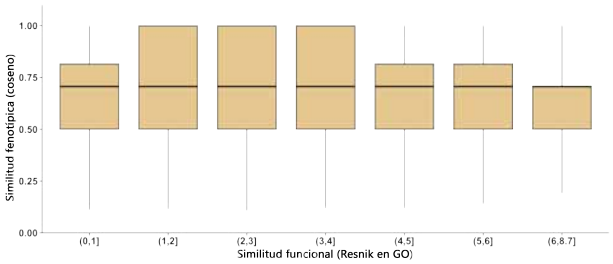
Figura 5.17: **Distribución de la similitud funcional en GO frente a la similitud fenotípica en CMPO filtrando los pares interactuantes en *iRef index*.** La tendencia es nula, aunque se hayan incrementado los valores de similitud semántica en CMPO.

5.6.4. De la función al fenotipo: Resultados para las métricas de similitud vectorial

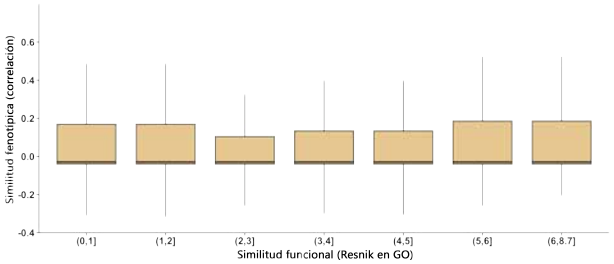
Esta ausencia de correlación entre la función y el fenotipo se ha observado también para otras similitudes fenotípicas distintas de la similitud fenotípica en CMPO. Con esto queda demostrado que no se trata de un efecto debido al tipo de métrica de similitud utilizada, sino que realmente no existe la tendencia que en principio se podría esperar.



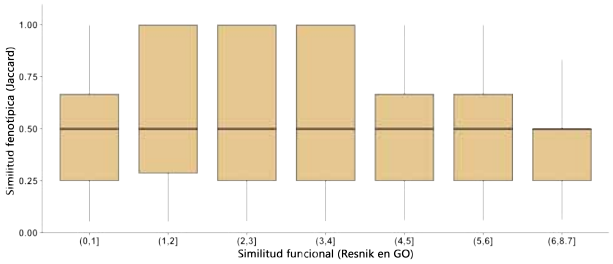
(a) Similitud funcional en GO frente a la similitud fenotípica TF-IDF.



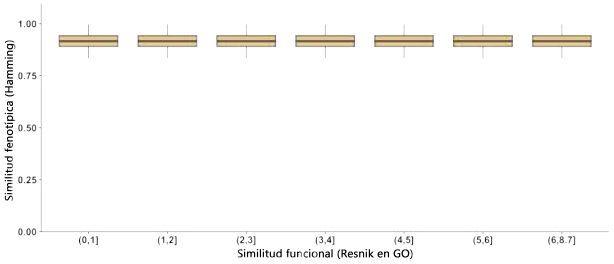
(b) Similitud funcional en GO frente a la similitud fenotípica de cosenos.



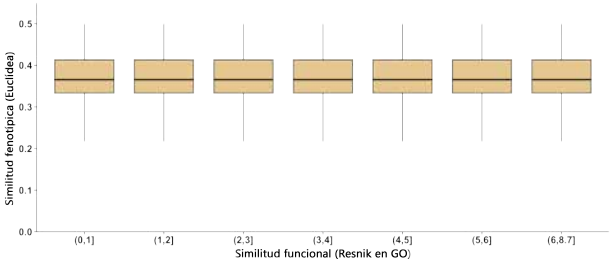
(c) Similitud funcional en GO frente a la similitud fenotípica usando correlación.



(d) Similitud funcional en GO frente a la similitud fenotípica Jaccard.



(e) Similitud funcional en GO frente a la similitud fenotípica Hamming.



(f) Similitud funcional en GO frente a la similitud fenotípica Euclídea.

Figura 5.18: **Similitud funcional en GO frente a la similitud fenotípica para distintas métricas.** No se observa tendencia alguna para las similitudes vectoriales, puede verse que para todas ellas la mediana se mantiene en una posición constante.

5.7. Estudio de la influencia de la alta dimensionalidad del espacio fenotípico

Hasta ahora, los resultados no son muy prometedores. Para descartar que se trate de un problema con las altas dimensiones de la matriz, se aplicó PCA a la matriz gen-fenotipo original. En esta nueva matriz (Figura 5.19), compuesta por números reales, se computaron algunas similitudes vectoriales, cuyos resultados ya se han mostrado en el dendrograma de la Figura 5.6, dentro de la sección 5.3.

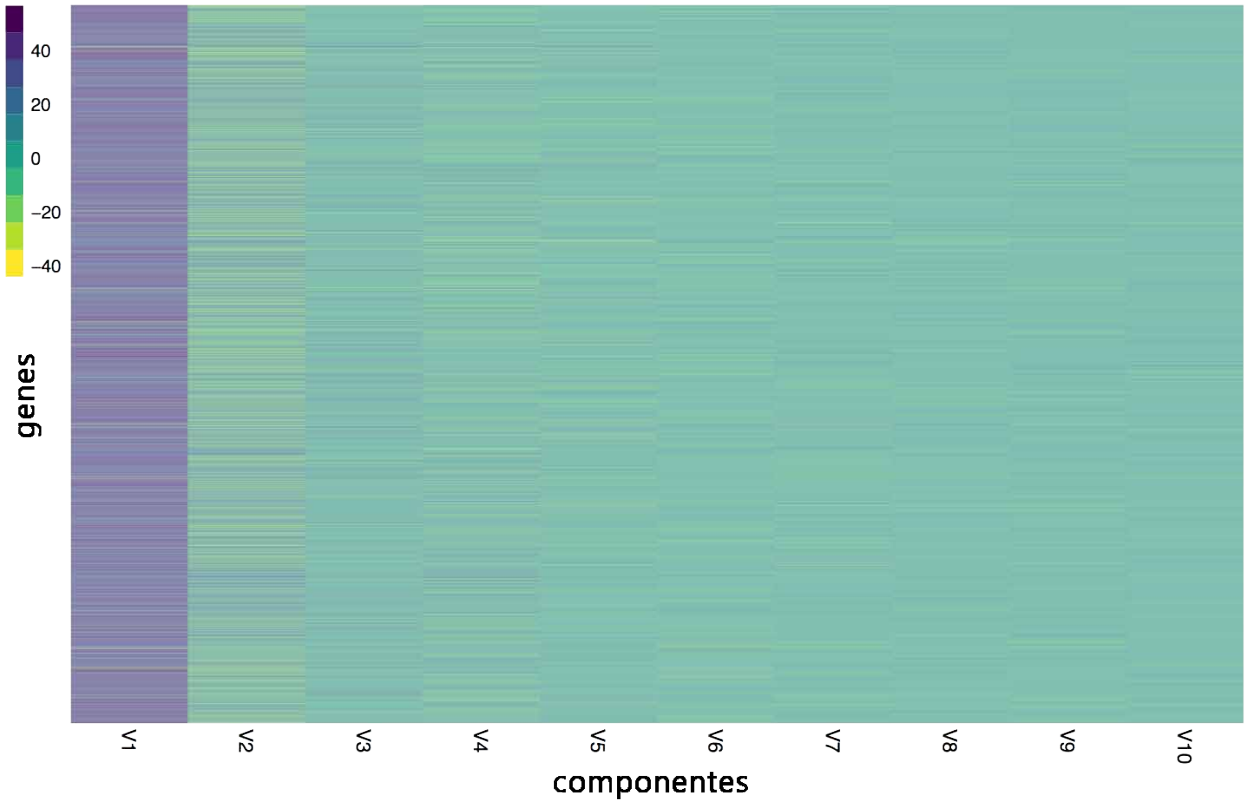


Figura 5.19: **Representación de la matriz gen-fenotipo reducida con PCA.** Las filas son los genes y las columnas las componentes, que serán combinaciones lineales de los 36 fenotipos originales. Se ajustó el modelo para obtener 10 componentes, que explican el 93.6% de la varianza.

A modo de ejemplo, se representaron los diagramas de cajas para la similitud semántica en GO y la fenotípica usando la similitud de cosenos (en ambas direcciones: similitud fenotípica vs. funcional, Figura 5.20; y similitud funcional vs. fenotípica, Figura 5.21).

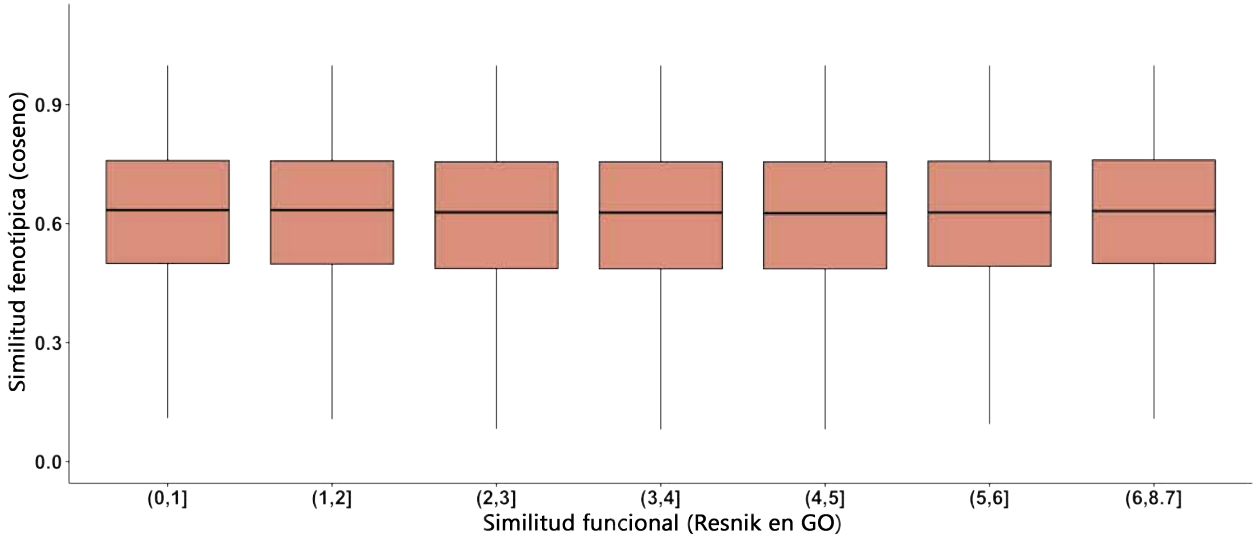


Figura 5.20: **Similitud funcional frente al coseno para la similitud fenotípica entre pares de genes.** La similitud fenotípica se aplicó a la matriz reducida mediante PCA.

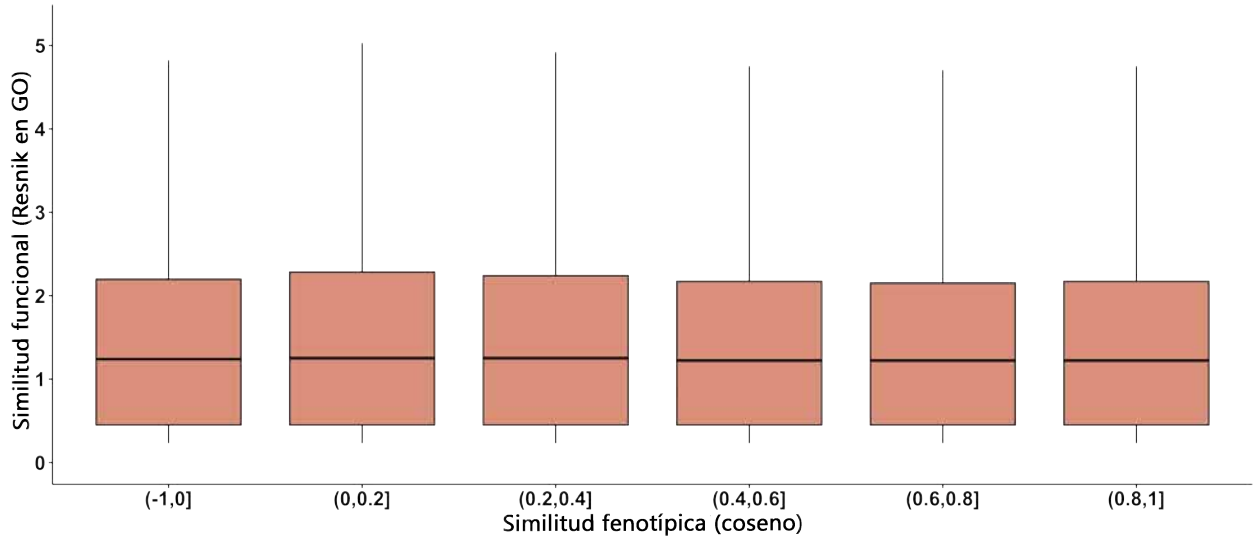


Figura 5.21: **Similitud fenotípica del coseno frente a similitud funcional entre pares de genes.** La similitud fenotípica se aplicó a la matriz reducida mediante PCA.

De nuevo, no se aprecia tendencia alguna, lo que confirma que no se trata de un problema con las dimensiones de los perfiles fenotípicos.

5.8. Análisis de los fenotipos a nivel individual

Las dos aproximaciones que se han visto hasta ahora basadas en la comparación de los perfiles fenotípicos –selección del término fenotípico más informativo por un lado y del perfil completo por otro–, no dan ninguna indicación de una relación significativa entre el fenotipo y la función de los genes. Este resultado es contrario a la intuición ya que la premisa de la mayoría de experimentos es que los genes con la misma función biológica dan lugar al mismo fenotipo descriptivo de una pérdida de función o mismo perfil fenotípico.

Para comprobar si esta premisa se mantiene en todos los experimentos distintos analizados, se calculó la similitud semántica media en GO para todos los pares de genes que muestran un fenotipo dado. Una vez hecho esto, se comparó esta media con la que se obtiene al aleatorizar 100 veces las asociaciones entre genes y fenotipos, aunque manteniendo invariable el número de aristas por fenotipo. Los resultados se muestran en la Figura 5.22.

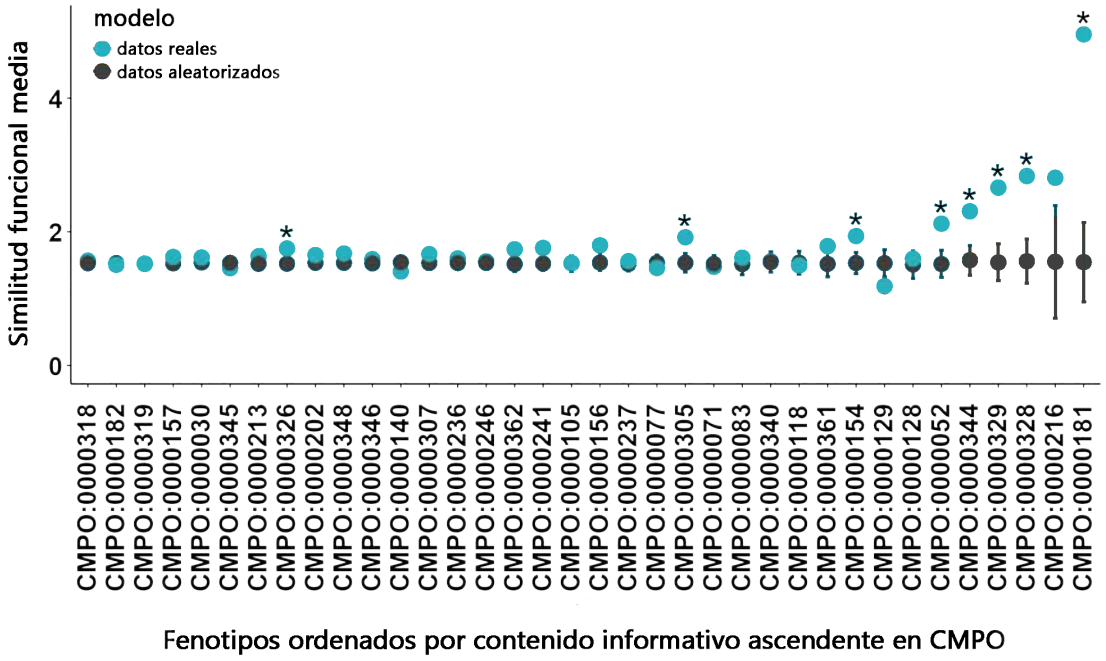


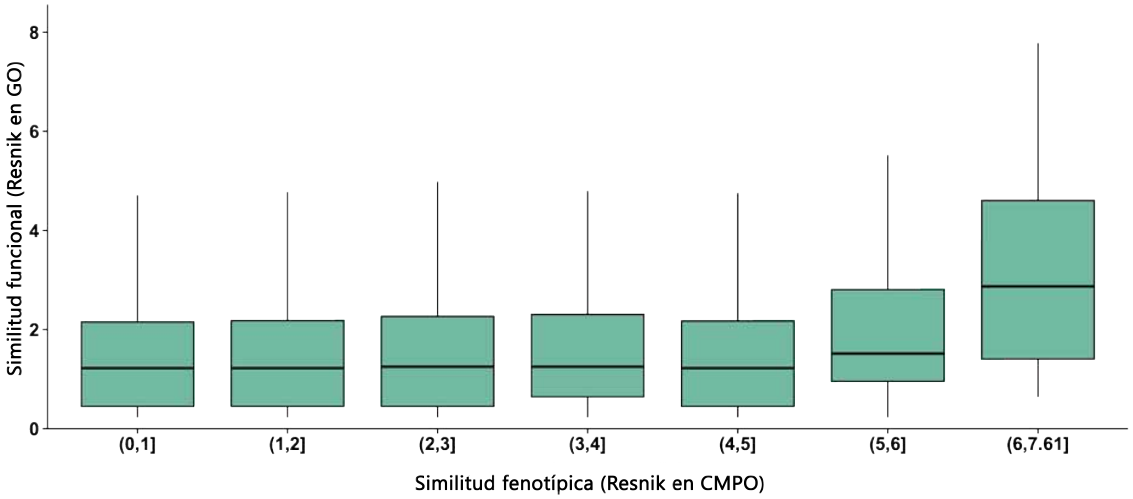
Figura 5.22: **Similitud semántica media en GO entre genes que comparten un fenotipo dado.** En azul se presenta la similitud semántica media en GO de todos los pares de genes que comparten un fenotipo. El modelo nulo está representado por la aleatorización de las relaciones entre genes y fenotipos (de color negro la media y las desviaciones típicas de las 100 aleatorizaciones). Los fenotipos con genes que tienen una similitud funcional alta (corrección FDR para el p-valor $\leqslant$0.01) están marcados con un asterisco. Los fenotipos están ordenados en el eje X de forma ascendente en cuanto a su contenido informativo en CMPO, de forma que los fenotipos más informativos ocupan la parte derecha. Las descripciones textuales asociadas a los términos CMPO del eje X pueden consultarse en la Tabla 4.1.

Un total de 8 sobre 36 fenotipos (aproximadamente el 25 %) dieron una señal significativa estadísticamente (p-valor ajustado mediante FDR$\leqslant$0.01) por tener su similitud funcional entre pares de genes por encima de la obtenida al aleatorizar. Más de la mitad de ellos corresponde a términos CMPO con alto contenido informativo, lo que indica que sólo los fenotipos más específicos tienden a asociarse con anotaciones funcionales en GO con alta similitud.

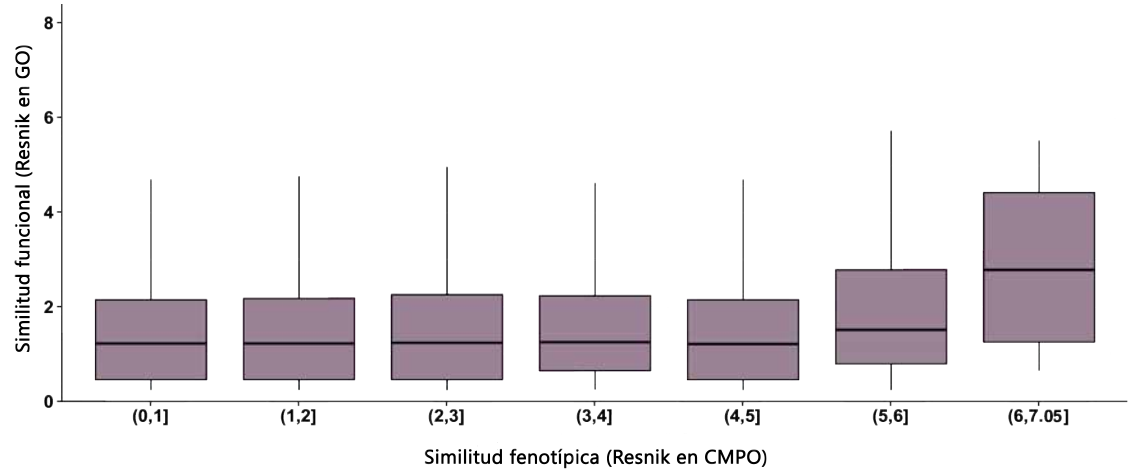
Estos resultados son más esperables, aunque la única lectura práctica que se deriva de la conversión automática de anotaciones fenotípicas a anotaciones funcionales es que sólo los fenotipos con alta similitud semántica son indicativos de función celular similar.

5.8.1. Estudio de la influencia de los fenotipos significativos en la tendencia ascendente entre la similitud fenotípica en CMPO y la funcional en GO

En la Figura 5.22, se observan 8 fenotipos con una señal estadísticamente significativa. Estos fenotipos podrían ser los únicos responsables de la débil tendencia observada en la Figura 5.8 (incluida de nuevo en la Figura 5.23a, en verde).



(a) Distribución de la similitud fenotípica en CMPO frente a la similitud funcional en GO para todos los pares de genes.



(b) Distribución de la similitud fenotípica en CMPO frente a la similitud funcional en GO sin tener en cuenta los fenotipos significativos.

Figura 5.23: **Distribución de la similitud fenotípica en CMPO frente a la similitud funcional en GO con y sin fenotipos significativos.** Para todos los rangos de similitud fenotípica en CMPO se obtienen distribuciones muy similares de similitud funcional usando Resnik en GO. Para facilitar la comparación se ha incluido de nuevo el diagrama de cajas original de la Figura 5.8.

Para descartar que los fenotipos significativos fuesen los únicos causantes de la tendencia, se excluyó temporalmente este grupo de fenotipos significativos de la matriz gen-fenotipo. La hipótesis es que, si estos fenotipos influyen de manera significativa en la tendencia, al obviarlos, la tendencia debería desaparecer para dar paso a una ausencia total de correlación. Sin embargo, en los resultados vemos que el diagrama de cajas se mantiene prácticamente invariable (Figura 5.23, morado), indicando que la tendencia que se observa no se debe únicamente al efecto de estos fenotipos altamente específicos.

5.9. Agrupamiento funcional de genes guiado por las anotaciones

Los resultados obtenidos hasta ahora sugieren que la estructura de GO no refleja adecuadamente las relaciones funcionales que subyacen a la similitud fenotípica en los experimentos utilizados. Recientemente se ha propuesto una forma alternativa de organizar los términos GO (Glass and Girvan, 2015) que consiste en agrupar términos basándose en las anotaciones compartidas entre ellos (Figura 5.24).

5.9.1. Proyección de las funciones celulares guiada por los genes

Para agrupar los términos GO basándose en las anotaciones de los genes, se genera un grafo de términos aplicando una proyección bipartita a la ontología (Figura 5.24a)), es decir, uniendo los términos en función de los genes anotados en ellos (Figura 5.24b).

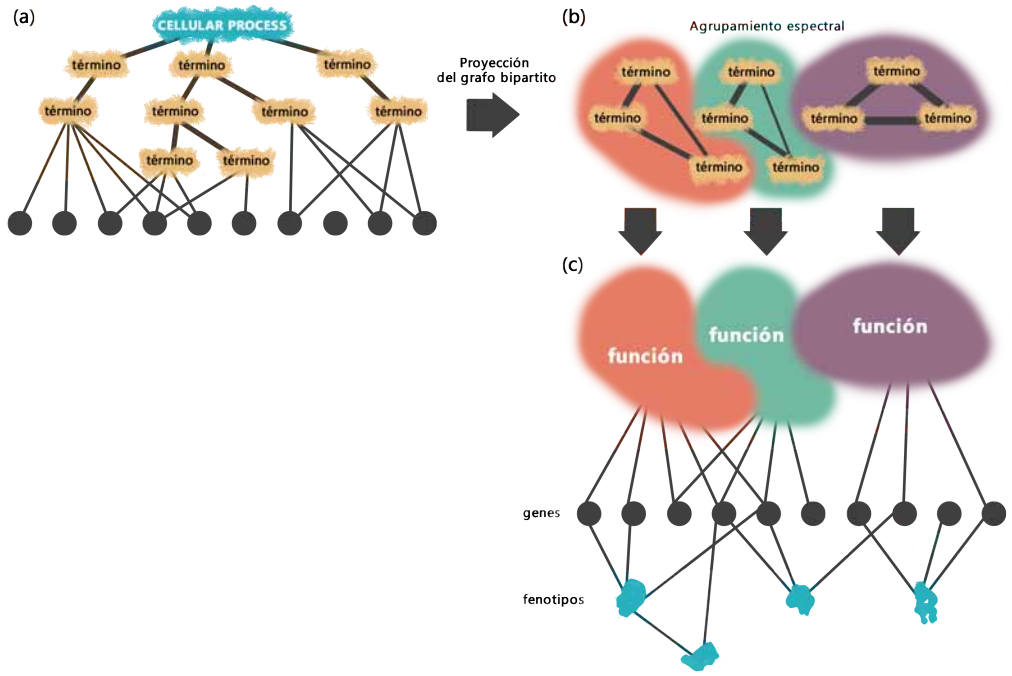


Figura 5.24: **Definición de funciones celulares guiada por las anotaciones. (a).** Los genes (círculos) se anotan a los términos (amarillo) dentro de la rama de procesos celulares. **(b).** Tras la proyección bipartita del grafo, los enlaces entre términos se pesan según el número de genes que comparten (ancho de línea). Luego, los términos se agrupan usando agrupamiento espectral. **(c).** Los clústeres de términos funcionales (nubes coloreadas en rojo, verde y morado) se enlazan con los fenotipos (azul) a través de los genes compartidos.

En la proyección (Figura 5.24b), dos términos son más similares cuanto más genes compartan en el grafo original (Figura 5.24a), es decir, cuanto más genes tengan anotados ambos en común. En la Figura 5.24b se representa la fortaleza de dicha unión con la anchura del trazo entre los términos.

Las funciones biológicas pueden entonces definirse como grupos de términos similares si se les aplica un algoritmo de clustering al grafo de términos. En este caso, el algoritmo usado es el agrupamiento espectral, explicado en las secciones 1.1.11.2 y 4.5.

Una vez aplicado el algoritmo de agrupamiento, las funciones pueden verse como un conjunto de términos co-ocurrentes. Estas nuevas funciones pueden vincularse a los fenotipos mediante las anotaciones de sus genes (Figura 5.24c).

La hipótesis ahora es si esta aproximación permitiría encontrar una relación –más fuerte que la obtenida hasta ahora– entre funciones y fenotipos. Para probar dicha hipótesis, se agruparon en primer lugar los 7470 términos de la rama de GO en estudio, *«cellular process»*, en un total de 140 clústeres.

5.9.2. Agrupamiento de las funciones biológicas basado en la red bipartita

Con el objetivo de decidir la granularidad de las funciones –o el nivel de especificidad–, el número de clústeres se fijó primero en 100. Sin embargo, tras esta partición se obtuvieron dos agrupaciones muy grandes (5050 y 817 términos) y el resto de menor tamaño. Como el nivel de granularidad presentaba una distribución muy dispar, se dividieron de nuevo los dos clústeres de mayor tamaño. Un criterio habitual para decidir el número de clústeres en el agrupamiento espectral suele ser encontrar el punto de mayor diferencia entre los autovalores de la matriz laplaciana.

En el caso del cluster de mayor tamaño –con 5050 términos funcionales–, se obtuvieron puntos de corte en 1, 3 y 33 clústeres (intervalos indicados en la Figura 5.25), por lo que se eligió 33 para obtener la granularidad deseada.

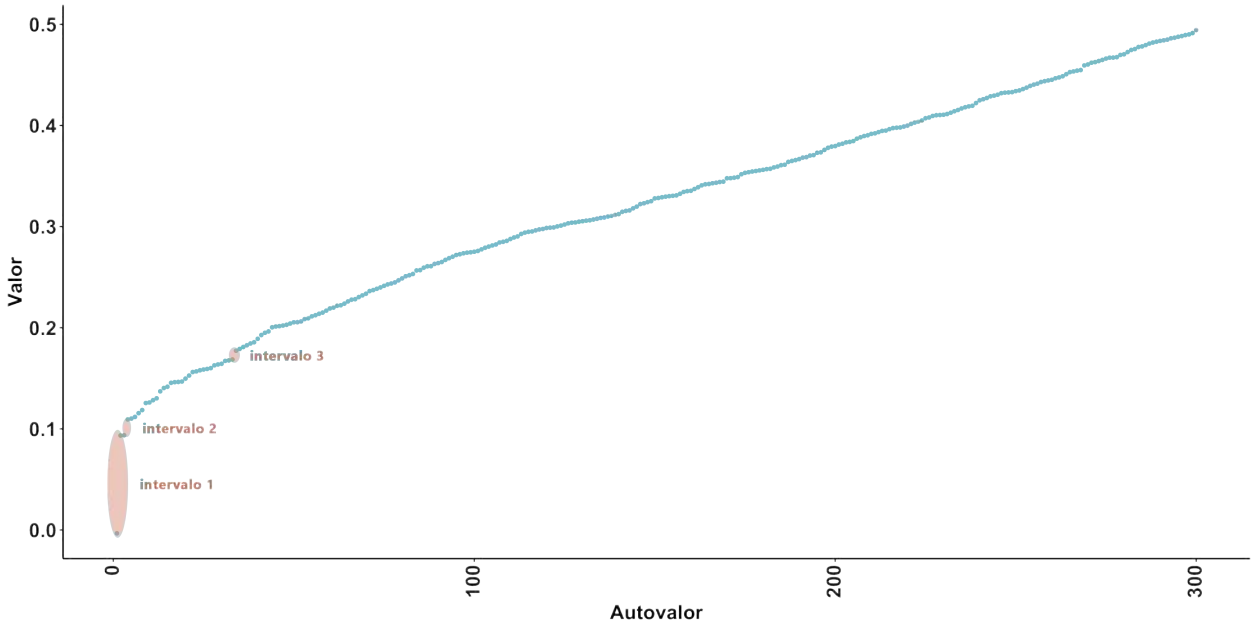


Figura 5.25: **Autovalores para el primer cluster de términos GO.** Aunque puede haber más de un punto de selección del número de clústeres, se seleccionaron 33 para obtener la granularidad deseada.

En el segundo cluster –con 817 términos GO– se obtuvo la distribución de autovalores de la Figura 5.26, con puntos de corte en 1, 9 y 4 (ordenados de mayor a menor según la distancia entre autovalores). En este caso, el número de grupos seleccionado fue de 9.

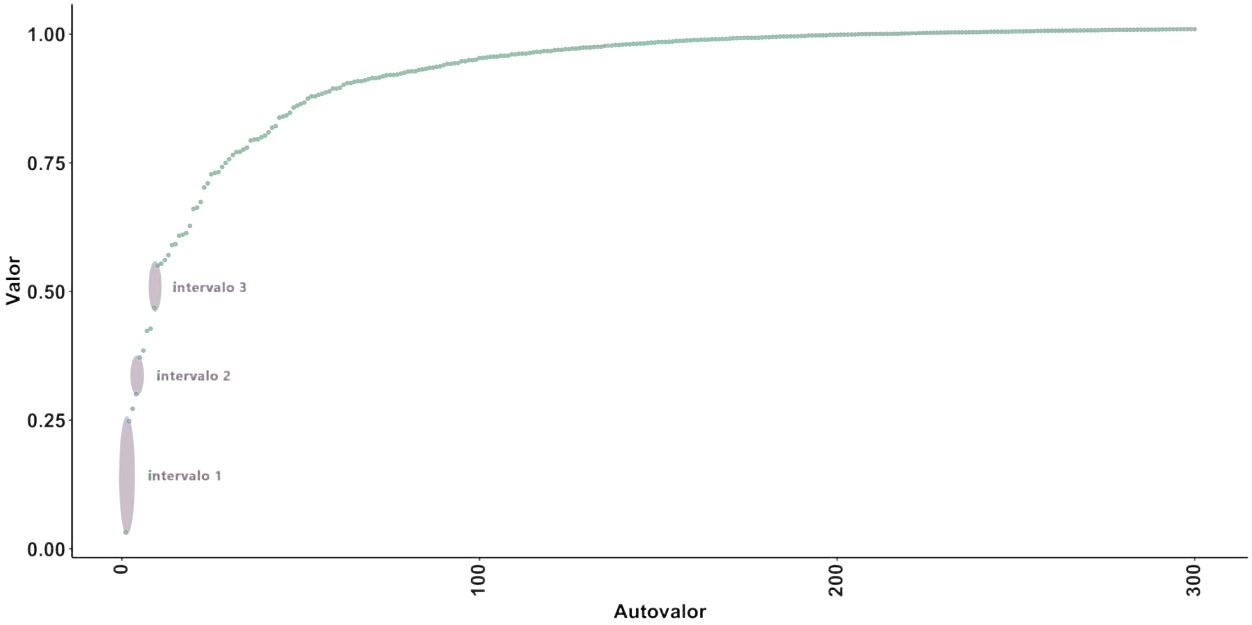


Figura 5.26: **Autovalores para el segundo cluster de términos GO.** Se seleccionaron en este caso 9 clústeres por ser el autovalor que presenta mayor diferencia con el anterior (a excepción del valor 1).

En total, con la partición inicial –98 grupos exceptuando los dos de mayor tamaño– y los dos grupos mayores –de 33 y 9 clústeres– se obtuvo un total de 140 agrupaciones. Estos 140 clústeres representan grupos de términos co-ocurrentes en cuanto a la anotación de los genes, o expresado de otro modo, conforman nuevas definiciones de función integrando varios términos de la ontología.

5.9.3. Validación de las relaciones fenotipo-función

Una forma de evaluar si esta nueva definición de función captura la similitud fenotípica consiste en calcular la similitud semántica de Resnik entre los términos de CMPO asociados –mediante genes– a cada clúster. Como modelo nulo, se permutaron 1000 veces las asociaciones entre términos y clústeres. El resultado se muestra en la Figura 5.27.

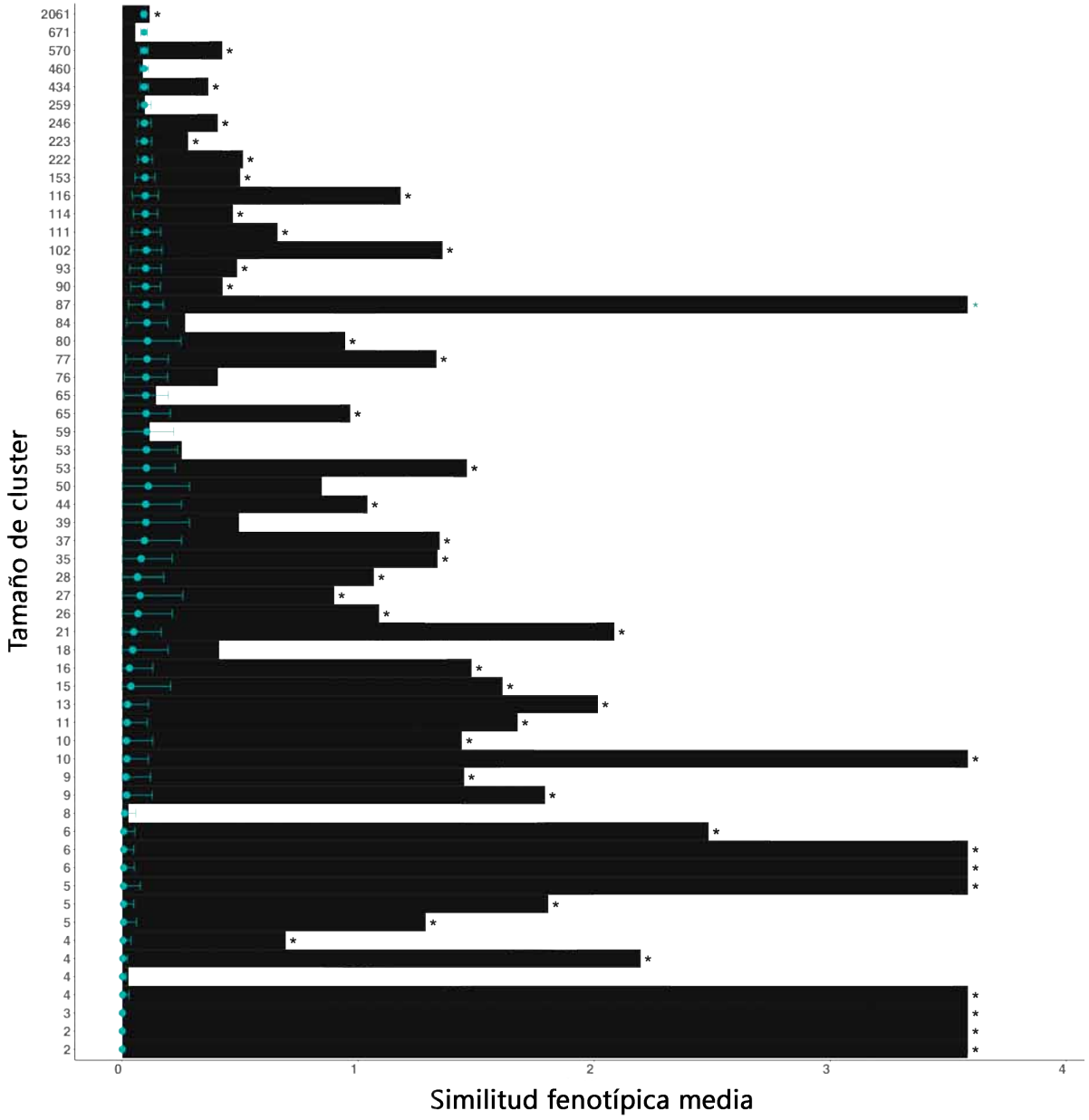


Figura 5.27: **Similitud fenotípica media en los clústeres de términos GO.** Cada barra simboliza la similitud fenotípica media de los términos CMPO asociados a los clústeres, que se ordenan en el eje Y por el número de términos que contienen. La media y desviación típica de la asignación aleatoria de términos a clústeres se representan en azul y las barras significativas se representan con un asterisco.

Sin tener en cuenta los clústeres que no están unidos a fenotipos, el 77 % (45/58) de los grupos funcionales tienen similitud fenotípica alta que no puede ser explicada por asignaciones al azar de términos GO a clústeres (con p-valores corregidos FDR $\leqslant$0.01, Figura 5.27).

Por lo tanto, se puede concluir que las funciones celulares generadas a partir de las anotaciones compartidas se asociaron con la similitud fenotípica.

5.10. Agrupamiento fenotípico de genes guiado por las anotaciones

Para comprobar la afirmación opuesta, es decir, si fenotipos similares reflejan funciones similares, se definió un grafo de términos fenotípicos de forma análoga pero en la ontología CMPO.

Esta vez, dado que el número de términos de CMPO es 361, el número de agrupaciones ha de ser menor. Para definir el número de clústeres, de nuevo se usaron los autovalores de la matriz laplaciana del agrupamiento espectral. Los puntos de corte obtenidos ahora son 3, 2, 13 y 10 (Figura 5.28). Se agrupó la ontología CMPO en 13 clústeres. El criterio aquí fue seleccionar un número de agrupaciones adecuado con más de un término para que tuviera sentido la agrupación guiada por anotaciones.

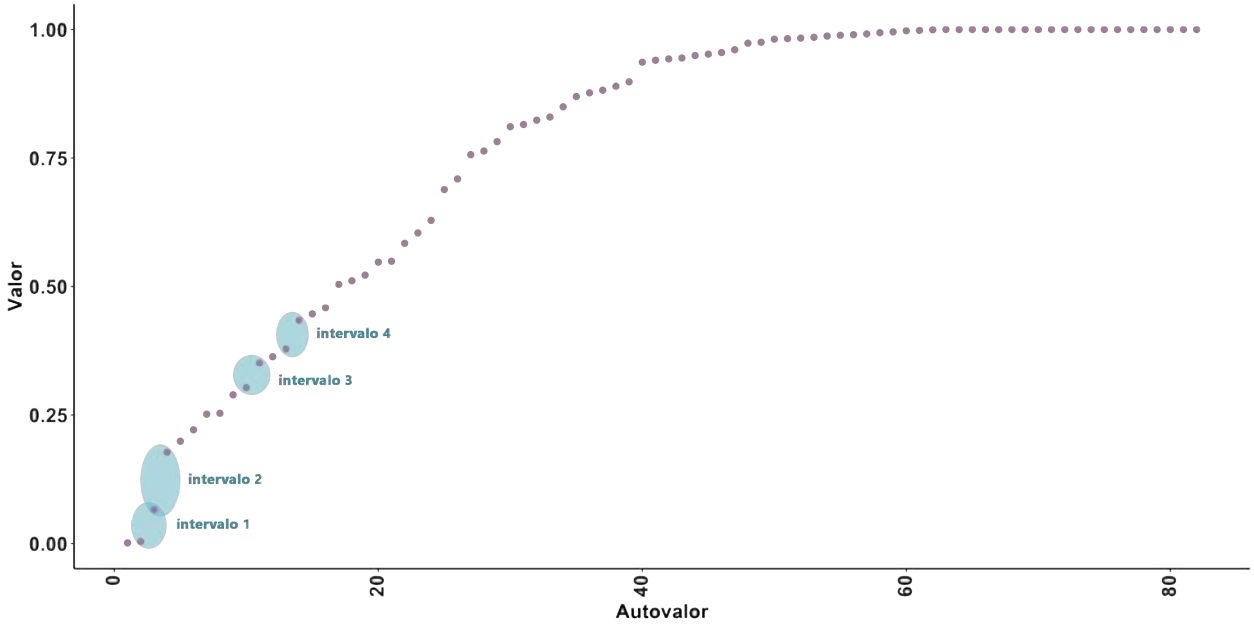


Figura 5.28: **Autovalores para la agrupación de términos CMPO.** La selección de 13 clústeres viene dada por el hecho de que un número mayor aumentaba el número de clústeres conteniendo un único término.

La visión ahora es opuesta: cada uno de los clústeres fenotípicos puede verse como un fenotipo caracterizado por una firma de descripciones fenotípicas co-ocurrentes. Al igual que en el caso contrario, se calculó la similitud semántica de Resnik entre términos GO dentro de los clústeres. De nuevo, excepto para los clústeres que no tienen anotaciones GO, se observó que la similitud funcional era más alta en los clústeres fenotípicos de lo que puede ser explicado por

asignaciones aleatorias de fenotipos a clústeres (Figura 5.29). Esto indica que esta definición de fenotipo sí permite recuperar la similitud funcional en GO.

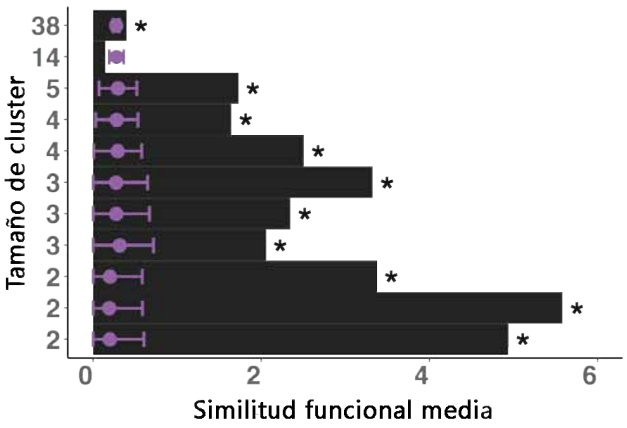


Figura 5.29: **Similitud funcional media en los clústeres de términos CMPO.** Las 1000 aleatorizaciones de las asignaciones de los términos a los clústeres se representan en morado. Los clústeres se ordenan por tamaño, es decir, número de términos.

Por todo lo expuesto, las funciones definidas como grupos de términos GO que comparten genes tienden a corresponderse con los términos CMPO mejor que las funciones definidas por términos GO individuales. En la otra dirección, los fenotipos definidos por grupos de términos CMPO asociados por genes se corresponden mejor con términos GO que los fenotipos definidos en términos individuales de CMPO.

Aunque los detalles de cómo los fenotipos y las funciones se definen están sujetos a cambios, tanto en GO como en CMPO, la asociación entre fenotipos y funciones es robusta, ya que sólo depende de los genes anotados.

I haven't failed. I've just found 10000 ways that won't work.

Thomas A. Edison (1847 – 1931)

6

Síntesis y líneas futuras

El principal reto de la genómica funcional es obtener las funciones de cada uno de los genes. Gracias a la gran cantidad de anotaciones fenotípicas procedentes de los experimentos de alto rendimiento algunas de esas funciones se han determinado, pero aún sigue siendo una fuente de información poco explotada.

En este trabajo se ha explorado cómo los fenotipos se relacionan con las funciones de los genes, con la esperanza de que se pudieran extraer algunos métodos que fuesen usados en la conversión automática de anotaciones fenotípicas en anotaciones funcionales. Para ello, se han usado fenotipos celulares procedentes de experimentos de silenciamiento génico usando RNAi en células humanas que han sido anotados en términos CMPO para estudiar cómo se relacionan esos fenotipos con las funciones de GO.

Con este objetivo, la primera pregunta a abordar fue cómo medir de forma adecuada la similitud fenotípica entre pares de genes. Esa similitud fenotípica se espera que esté correlacionada con la similitud funcional para que refleje adecuadamente la relación entre las anotaciones fenotípicas y funcionales. En este sentido, tras analizar distintas métricas, se encontró que las medidas basadas en el contenido informativo presentan una mayor correlación con las interacciones entre proteínas, mientras que las que se basan en similitud vectorial no parecieron ser muy útiles para este propósito. Entre esas medidas basadas en el contenido informativo destacó la similitud semántica de Resnik, que hace uso de la estructura jerárquica entre los fenotipos en CMPO.

Sin embargo, a pesar de que las medidas de similitud semántica son prometedoras, en nuestro estudio sólo los fenotipos con alta similitud en CMPO se asociaron con anotaciones funcionales en procesos celulares de GO. Para vincular las partes de los dos espacios que se correlacionan, podría definirse un valor umbral de similitud semántica a partir del cual sería fiable asociar fenotipos con funciones. Este umbral es especialmente complicado de definir cuando se trata de ontologías, ya que son susceptibles de expandirse y eso provoca cambios en los valores de similitud semántica. Además, sólo una pequeña fracción de los genes se podrían anotar con funciones celulares de esta forma, pues para la mayoría de casos no existe relación entre la similitud semántica y fenotípica.

Dado que los perfiles fenotípicos no parecían ser una forma adecuada de representación para obtener una correlación con la mayoría de funciones en GO, se planteó un cambio en la aproximación que aprovechara mejor la información de partida. Definir las funciones celulares como grupos de términos GO co-ocurrentes nos permitió recuperar la relación entre la similitud fenotípica y la funcional. También al contrario: definir los fenotipos como grupos de términos CMPO co-ocurrentes nos permitió recuperar la relación entre los fenotipos y las funciones similares en GO.

Con estas nuevas definiciones de fenotipo y función como entidades complejas y conexas, se observó que las funciones celulares similares llevan a fenotipos similares; y que fenotipos similares llevan a funciones similares. Estos resultados refrendan la observación de (Glass and Girvan, 2015), que no encuentra la asociación directa entre firmas genéticas del cáncer y ramas de la ontología GO, pero sí con agrupaciones de términos de distintas ramas según las anotaciones que comparten.

De alguna forma, esta definición de función biológica ya la usan algunos algoritmos de priorización de genes basados en redes. Este es el caso de *FUN-L* (Lees et al., 2015) y *GeneMANIA* (Mostafavi et al., 2008), que parten de una lista de genes relacionados funcionalmente. Esta definición de función contribuye, de acuerdo a los resultados aquí obtenidos, a su éxito al enriquecer genes candidatos en los fenotipos deseados.

Todo lo observado aquí tiene varias implicaciones prácticas. La primera es aplicable en el ámbito experimental: la forma habitual de agrupar los perfiles fenotípicos –mediante métricas vectoriales– no parece ser la mejor de las aproximaciones para predecir la función de los genes. Se ha demostrado en este trabajo que estos tipos de métricas presentan poca correlación con la similitud funcional. Una posible mejora sería tener en cuenta el contenido informativo de los fenotipos.

A la luz de los resultados obtenidos, una aproximación más coherente, en vez de agrupar los genes, consiste en agrupar los fenotipos según los genes anotados en ellos. Una vez hecho esto, se podría calcular el enriquecimiento en términos funcionales de esos clústeres de fenotipos. Al final, los genes asociados a un clúster de términos CMPO pueden anotarse también con los términos GO con los que se relaciona. Esta aplicación es relevante para las tareas de anotación

tanto en los procesos de curación de datos ya existentes como en los análisis de un experimento de RNAi concreto donde se estudien distintos fenotipos.

La segunda implicación se refiere a la integración de información fenotípica con otros datos biológicos. Varios métodos de selección de genes candidatos se basan en la combinación de distintas fuentes de información para aumentar su precisión y ampliar el alcance de las asociaciones funcionales entre genes humanos. De momento, la información fenotípica no se usa en esos esquemas de integración. Para incorporarla, una opción consiste en aplicar métodos de aprendizaje automático supervisado que podrían anotar funcionalmente los genes a partir de anotaciones fenotípicas.

Para un algoritmo de aprendizaje supervisado (como SVM, por ejemplo) es necesario partir de un conjunto de entrenamiento de calidad que garantice que el resultado del algoritmo sea correcto. Partiendo de un gen y su fenotipo asociado, la salida del algoritmo sería la predicción de la función. En este caso, el conjunto de entrenamiento estaría formado por asociaciones del tipo fenotipo-función previamente conocidas y fiables. Sin embargo, como la relación entre los fenotipos y las funciones no parecen estar muy claramente definidas, el conjunto de entrenamiento no sería el adecuado.

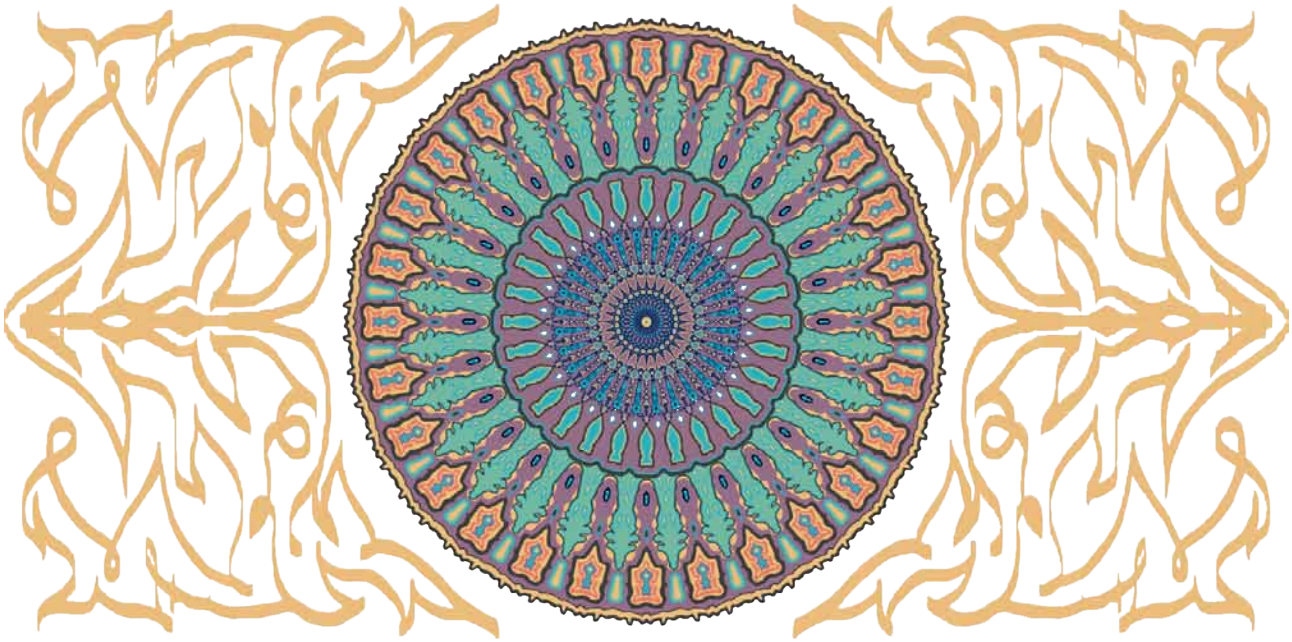
Siendo consistente con los resultados que aquí se presentan, usar las métricas vectoriales habituales como medidas de similitud fenotípica lleva a un bajo rendimiento a la hora de encontrar genes relacionados funcionalmente. De acuerdo con nuestros resultados, si se reemplazan los fenotipos individuales por clústeres de términos CMPO, se pueden obtener mejores resultados en la integración. De forma análoga, si se consideran las enfermedades como fenotipos, la similitud funcional derivada de las agrupaciones –guiadas por las anotaciones en GO– podrían ser más útiles para predecir genes relacionados con enfermedades que la similitud funcional basada únicamente en la similitud semántica.

If you torture data enough, nature will always confess.

R. H. Coase (1910 – 2013)

7

Conclusiones



7.1. Conclusiones

Tras todo lo expuesto en los capítulos anteriores, se enumeran a continuación las conclusiones de esta Tesis Doctoral:

❶ La comparación de distintas métricas de similitud entre perfiles fenotípicos determina que Resnik sobre la ontología CMPO es la medida que mejor representa la relación funcional entre pares de genes.

❷ A nivel celular, no es posible derivar las anotaciones funcionales de los genes en GO a partir de sus perfiles fenotípicos experimentales ni viceversa.

❸ La agrupación de términos ontológicos guiada por las anotaciones de los genes proporciona una estructura alternativa y complementaria a las ontologías en la que convergen los espacios fenotípico y funcional.

7.2. Conclusions

Several conclusions may be drawn from this Doctoral Thesis:

❶ The comparison of different similarity measures between phenotypic profiles determines that Resnik on the CMPO ontology is the measure that best represents the functional relationship between gene pairs.

❷ At a cellular level, it is not possible to derive the GO functional annotations of genes from the experimental phenotypic profiles or vice versa.

❸ Annotation-driven clustering of the ontological terms provides an alternative and complementary structure to the ontologies in which the phenotypic and functional spaces converge.

Bibliografía

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29.

Bakal, C., Church, G., and Perrimon, N. (2007). Regulating Cell Morphology. *Science*, 316(June):1753–1756.

Balestra, F., Strnad, P., Flückiger, I., and Gönczy, P. (2013). Discovering regulators of centriole biogenesis through siRNA-based functional genomics in human cells. *Developmental Cell*, 25(6):555–571.

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLOS Computational Biology*, 4(10).

Birmingham, A., Selfors, L. M., Forster, T., Wrobel, D., Kennedy, J., Shanks, E., Santoyo-lopez, J., Dunican, D. J., Kelleher, D., Smith, Q., Beijersbergen, R. L., and Shamu, C. E. (2009). Interference Screens. *Nature Methods*, 6(8):569–575.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer.

Blohm, P., Frishman, G., Smialowski, P., Goebels, F., Wachinger, B., Ruepp, A., and Frishman, D. (2014). Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Research*, 42(D1):396–400.

Bodenreider, O. and Stevens, R. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7(3):256–274.

Boettcher, M. and McManus, M. T. (2015). Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. *Molecular Cell*, 58(4):575–585.

Boutros, M. and Ahringer, J. (2008). The art and design of genetic screens: RNA interference. *Nature Reviews. Genetics*, 9(7):554–566.

Boutros, M., Brás, L. P., and Huber, W. (2006). Analysis of cell-based RNAi screens. *Genome Biology*, 7(7):R66.

Boutros, M., Heigwer, F., and Laufer, C. (2015). Microscopy-Based High-Content Screening. *Cell*, 163(6):1314–1325.

Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *International Journal of Molecular Sciences*, 10(6):2763–2788.

Caicedo, J. C., Singh, S., and Carpenter, A. E. (2016). Applications in image-based profiling of perturbations. *Current Opinion in Biotechnology*, 39:134–142.

Camon, E. (2004). The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32(90001):262D–266D.

Campbell, A. E. and Bennett, D. (2016). Targeting protein function: the expanding toolkit for conditional disruption. *Biochemical Journal*, 473(17):2573–2589.

Conway, J. B. (2012). *A Course in Abstract Analysis*. American Mathematical Society, Providence, Rhode Island.

Côté, R., Jones, P., Apweiler, R., and Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics*, 7(1):97.

Degtyarenko, K., De matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(SUPPL. 1):344–350.

Dessimoz, C. and Škunca, N. (2017). *The Gene Ontology Handbook*, volume 1446 of *Methods in Molecular Biology*. Springer New York, New York, NY.

Dessimoz, C. and Walker, J. M. (2016). *The Gene Ontology Handbook IN Series Editor*. Humana Press, Springer Open.

du Plessis, L., Skunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735.

Dunham, W. H., Mullin, M., and Gingras, A. C. (2012). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics*, 12(10):1576–1590.

Echeverri, C. J., Beachy, P. A., Baum, B., Boutros, M., Buchholz, F., Chanda, S. K., Downward, J., Ellenberg, J., Fraser, A. G., Hacohen, N., Hahn, W. C., Jackson, A. L., Kiger, A., Linsley, P. S., Lum, L., Ma, Y., Mathey-Prevot, B., Root, D. E., Sabatini, D. M., Taipale, J., Perrimon, N., and Bernards, R. (2006). Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nature Methods*, 3(10):777–779.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., Matthews, L., May, B., Milacic, M., Rothfels, K., Shamovsky, V., Webber, M., Weiser, J., Williams, M., Wu, G., Stein, L., Hermjakob, H., and D'Eustachio, P. (2016). The Reactome pathway Knowledgebase. *Nucleic Acids Research*, 44(D1):D481–D487.

Fang, H. and Gough, J. (2014). The 'dnet' approach promotes emerging research on cancer patient survival. *Genome Medicine*, 6:64.

Feng, Y., Mitchison, T. J., Bender, A., Young, D. W., and Tallarico, J. A. (2009). Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. *Nat Rev Drug Discov*, 8(7):567–578.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811.

Fuchs, F., Pau, G., Kranz, D., Sklyar, O., Budjan, C., Steinbrink, S., Horn, T., Pedal, A., Huber, W., and Boutros, M. (2010). Clustering phenotype populations by genome-wide RNAi and multiparametric imaging. *Molecular Systems Biology*, 6(370):370.

Gaj, T. (2014). ZFN, TALEN and CRISPR/Cas based methods for genome engineering. *Trends in Biotechnology*, 31(7):397–405.

Glass, K. and Girvan, M. (2015). Finding New Order in Biological Functions from the Network Structure of Gene Annotations. *PLOS Computational Biology*, 11(11):e1004565.

Granas, C. (2006). Identification of RAS-Mitogen-Activated Protein Kinase Signaling Pathway Modulators in an ERF1 Redistribution(R) Screen. *Journal of Biomolecular Screening*, 11(4):423–434.

Greenacre, M. and Primicerio, R. (2013). Hierarchical Cluster Analysis. *Multivariate Analysis of Ecological Data*, (December):87–99.

Groth, P., Weiss, B., Pohlenz, H.-D., and Leser, U. (2008). Mining phenotypes for gene function prediction. *BMC Bioinformatics*, 9(1):136.

Gunsalus, K. C., Yueh, W.-C., MacMenamin, P., and Piano, F. (2004). RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Research*, 32(Database issue):D406–D410.

Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8):967–973.

Gururaja, T. L., Goff, D., Kinoshita, T., Goldstein, E., Yung, S., McLaughlin, J., Pali, E., Huang, J., Singh, R., Daniel-Issakani, S., Hitoshi, Y., Cooper, R. D. G., and Payan, D. G. (2006). R-253 disrupts microtubule networks in multiple tumor cell lines. *Clinical Cancer Research*, 12(12):3831–3842.

Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M. (2012). Semantic similarity analysis of protein data: Assessment with biological features and issues. *Briefings in Bioinformatics*, 13(5):569–585.

Hancock, J. M. (2014). *Phenomics*. CRC Press.

Hannon, G. J. and Rossi, J. J. (2004). Unlocking the potential of the human genome with RNA interference. *Nature*, 431(7006):371–378.

Hennig, C. and Hausdorf, B. (2015). *prabclus: Functions for Clustering of Presence-Absence, Abundance and Multilocus Genetic Data*. R package version 2.2-6.

Hériché, J.-K., Lees, J. G., Morilla, I., Walter, T., Petrova, B., Roberti, M. J., Hossain, M. J., Adler, P., Fernandez, J. M., Krallinger, M., Haering, C. H., Vilo, J., Valencia, A., Ranea, J. A., Orengo, C., and Ellenberg, J. (2014). Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. *Molecular Biology of the Cell*, 25(16):2522–2536.

Hoehndorf, R., Harris, M. A., Herre, H., Rustici, G., and Gkoutos, G. V. (2012). Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology. *Bioinformatics*, 28(13):1783–1789.

Hoehndorf, R., Schofield, P. N., and Gkoutos, G. V. (2015). The role of ontologies in biological and biomedical research: A functional perspective. *Briefings in Bioinformatics*, 16(6):1069–1080.

Hu, P., Janga, S. C., Babu, M., Díaz-Mejía, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., Chandran, S., Christopoulos, C., Nazarians-Armavil, A., Nasseri, N. K., Musso, G., Ali, M., Nazemof, N., Eroukova, V., Golshani, A., Paccanaro, A., Greenblatt, J. F., Moreno-Hagelsieb, G., and Emili, A. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLOS Biology*, 7(4):e96.

Jackson, A. L. and Linsley, P. S. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery*, 9(1):57–67.

Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562.

Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proceedings of International Conference Research on Computational Linguistics*, cmp-lg/9709008(Rocling X):19–33.

Johannessen, C. M., Clemons, P. A., and Wagner, B. K. (2015). Integrating phenotypic small-molecule profiling and human genetics: The next phase in drug discovery. *Trends in Genetics*, 31(1):16–23.

Jolliffe, I. T. (2002). Principal Component Analysis, Second Edition. *Encyclopedia of Statistics in Behavioral Science*, 30(3):487.

Jupp, S., Malone, J., Burdett, T., Heriche, J.-K., Williams, E., Ellenberg, J., Parkinson, H., and Rustici, G. (2016). The cellular microscopy phenotype ontology. *Journal of Biomedical Semantics*, 7(1):28.

Kabán, A. (2012). Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, 22(2):375–385.

Kirsanova, C., Brazma, A., Rustici, G., and Sarkans, U. (2015). Cellular phenotype database: a repository for systems microscopy data. *Bioinformatics*, 31(16):2736–2740.

Kitano, H. (2002). Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664.

Landgraf, A. J., Lee, Y., and Oct, M. L. (1999). Dimensionality Reduction for Binary Data through the Projection of Natural Parameters. *arXiv*.

Laufer, C., Fischer, B., Billmann, M., Huber, W., and Boutros, M. (2013). Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. *Nature Methods*, 10(5):427–431.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A. G., and Marcotte, E. M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature Genetics*, 40(2):181–188.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. a. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews. Genetics*, 11(10):733–739.

Lees, J. G., Hériché, J. K., Morilla, I., Fernández, J. M., Adler, P., Krallinger, M., Vilo, J., Valencia, A., Ellenberg, J., Ranea, J. A., and Orengo, C. (2015). FUN-L: Gene prioritization for RNAi screens. *Bioinformatics*, 31(12):2052–2053.

Li, Z. (2003). Identification of Gap Junction Blockers Using Automated Fluorescence Microscopy Imaging. *Journal of Biomolecular Screening*, 8(5):489–499.

Liberali, P., Snijder, B., and Pelkmans, L. (2014). Single-cell and multivariate approaches in genetic perturbation screens. *Nature Reviews Genetics*, 16(1):18–32.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *Proceedings of ICML*, pages 296–304.

Ljosa, V., Caie, P. D., ter Horst, R., Sokolnicki, K. L., Jenkins, E. L., Daya, S., Roberts, M. E., Jones, T. R., Singh, S., Genovesio, A., Clemons, P. A., Carragher, N. O., and Carpenter, A. E. (2013). Comparison of Methods for Image-Based Profiling of Cellular Morphological Responses to Small-Molecule Treatment. *Journal of Biomolecular Screening*, 18(10):1321–1329.

Lock, J. G. and Strömblad, S. (2010). Systems microscopy: An emerging strategy for the life sciences. *Experimental Cell Research*, 316(8):1438–1444.

Loo, L.-H., Wu, L. F., and Altschuler, S. J. (2007). Image-based multivariate profiling of drug responses from single cells. *Nature Methods*, 4(5):445–453.

Lundholt, B. K. B. K., Scudder, K. M., and Pagliaro, L. (2003). A Simple Technique for Reducing Edge Effect in Cell-Based Assays. *Journal of Biomolecular Screening*, 8(5):566–570.

Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., and Nadon, R. (2006). Statistical practice in high-throughput screening data analysis. *Nature Biotechnology*, 24(2):167–175.

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, 27(February):1729–1736.

Mazandu, G. K. and Mulder, N. J. (2012). A topology-based metric for measuring term similarity in the gene ontology. *Advances in Bioinformatics*, 2012.

Mazandu, G. K. and Mulder, N. J. (2014). Information content-based gene ontology functional similarity measures: Which one to use for a given biological data type? *PLOS ONE*, 9(12):1–20.

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189.

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., D'Eustachio, P., and Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers*, 4(4):1180–1211.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 101(2):343–352.

Mistry, M. and Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327.

Moreau, Y. and Tranchevent, L.-C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(July):1–14.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biology*, 9 Suppl 1:S4.

Moudry, P., Lukas, C., Macurek, L., Neumann, B., Heriche, J.-K., Pepperkok, R., Ellenberg, J., Hodny, Z., Lukas, J., and Bartek, J. (2012). Nucleoporin NUP153 guards genome integrity by promoting nuclear import of 53BP1. *Cell Death and Differentiation*, 19(5):798–807.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E., and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2.

Nachtomy, O., Shavit, A., and Yakhini, Z. (2007). Gene expression and the concept of the phenotype. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 38(1):238–254.

Neumann, B., Walter, T., Hériché, J.-K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., Cetin, C., Sieckmann, F., Pau, G., Kabbe, R., Wünsche, A., Satagopam, V., Schmitz, M. H. A., Chapuis, C., Gerlich, D. W., Schneider, R., Eils, R., Huber, W., Peters, J.-M., Hyman, A. A., Durbin, R., Pepperkok, R., and Ellenberg, J. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727.

Newman, M. (2010). *Networks. An Introduction*.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., and Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(Web Server issue):W170–W173.

Oellrich, A., Collier, N., Groza, T., Rebholz-Schuhmann, D., Shah, N., Bodenreider, O., Boland, M. R., Georgiev, I., Liu, H., Livingston, K., Luna, A., Mallon, A. M., Manda, P., Robinson, P. N., Rustici, G., Simon, M., Wang, L., Winnenburg, R., and Dumontier, M. (2016). The digital revolution in phenotyping. *Briefings in Bioinformatics*, 17(5):819–830.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34.

Okayama, H. and Berg, P. (1982). High-efficiency cloning of full-length cDNA. *Molecular and Cellular Biology*, 2(2):161–170.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Van Roey, K., Cesareni, G., and Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, 42(D1):358–363.

Orlov, N., Shamir, L., Macura, T., Johnston, J., Eckley, D. M., and Goldberg, I. G. (2008). WND-CHARM: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693.

Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stümpflen, V., Mewes, H. W., Ruepp, A., and Frishman, D. (2005). The MIPS mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834.

Paulsen, R. D., Soni, D. V., Wollman, R., Hahn, A. T., Yee, M.-C., Guan, A., Hesley, J. A., Miller, S. C., Cromwell, E. F., Solow-Cordero, D. E., Meyer, T., and Cimprich, K. A. (2009). A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability. *Molecular Cell*, 35(2):228–239.

Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F., and Altschuler, S. J. (2004). Multidimensional drug profiling by automated microscopy. *Science (New York, N.Y.)*, 306(5699):1194–1198.

Pesquita, C., Faria, D., Bastos, H., Falcão, A. O., and Couto, F. M. (2007). *Evaluating GO-based semantic similarity measures*, volume 2007.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009). Semantic Similarity in Biomedical Ontologies. *PLOS Computational Biology*, 5(7):e1000443.

Qi, Y., Suhail, Y., Lin, Y.-y., Boeke, J. D., and Bader, J. S. (2008). Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Research*, 18(12):1991–2004.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.

Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). On the existence of obstinate results in vector space models. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10*, 1(Section 2):186–193.

Rajaram, S., Pavie, B., Wu, L. F., and Altschuler, S. J. (2012). PhenoRipper: software for rapidly profiling microscopy images. *Nature Methods*, 9(7):635–637.

Razick, S., Magklaras, G., and Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, 9:405.

Reisen, F., Zhang, X., Gabriel, D., and Selzer, P. (2013). Benchmarking of multivariate similarity measures for high-content screening fingerprints in phenotypic drug discovery. *Journal of Biomolecular Screening*, 18(10):1284–1297.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1 - IJCAI'95*, 1:6.

Richards, G. R., Smith, A. J., Parry, F., Platts, A., Chan, G. K., Leveridge, M., Kerby, J. E., and Simpson, P. B. (2006). A morphology- and kinetics-based cascade for human neural cell high content screening. *Assay Drug Dev Technol*, 4(2):143–152.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Rojas, A. M., Santamaria, A., Malik, R., Jensen, T. S., Körner, R., Morilla, I., de Juan, D., Krallinger, M., Hansen, D. A., Hoffmann, R., Lees, J., Reid, A., Yeats, C., Wehner, A., Elowe, S., Clegg, A. B., Brunak, S., Nigg, E. A., Orengo, C., Valencia, A., and Ranea, J. A. G. (2012). Uncovering the molecular machinery of the human spindle–an integration of wet and dry systems biology. *PLOS ONE*, 7(3):e31813.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–D451.

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302.

Schmidt, E. E., Pelz, O., Buhlmann, S., Kerr, G., Horn, T., and Boutros, M. (2013). GenomeRNAi: A database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Research*, 41(D1):1021–1026.

Sebastian, K., Robinson, P. N., and Mungall, C. J. (2017). Opposite-of-information improves similarity calculations in phenotype ontologies. *bioRxiv*.

Shalem, O., Sanjana, N. E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. *Nature Reviews. Genetics*, 16(5):299–311.

Shamir, L. (2011). Assessing the efficacy of low-level image content descriptors for computer-based fluorescence microscopy image analysis. *Journal of Microscopy*, 243(3):284–292.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston, J., and Goldberg, I. G. (2008). Wndchrm - an open source utility for biological image analysis. *Source Code for Biology and Medicine*, 3:13.

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 623–656.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, 3(88):88.

Simpson, J. C., Joggerst, B., Laketa, V., Verissimo, F., Cetin, C., Erfle, H., Bexiga, M. G., Singan, V. R., Hériché, J.-K., Neumann, B., Mateos, A., Blake, J., Bechtel, S., Benes, V., Wiemann, S., Ellenberg, J., and Pepperkok, R. (2012). Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. *Nature Cell Biology*, 14(7):764–774.

Slack, M. D., Martinez, E. D., Wu, L. F., and Altschuler, S. J. (2008). Characterizing heterogeneous cellular responses to perturbations. *Proceedings of the National Academy of Sciences of the United States of America*, 105(49):19306–19311.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., and Christopher, J. (2010). NIH Public Access. *Nature Biotechnology*, 25(11):1251–1255.

Smith, B., Williams, J., and Schulze-Kremer, S. (2003). The ontology of the gene ontology. *Annual Symposium Proceedings / AMIA Symposium*, 2003:609–613.

Snijder, B., Sacher, R., Rämö, P., Liberali, P., Mench, K., Wolfrum, N., Burleigh, L., Scott, C. C., Verheije, M. H., Mercer, J., Moese, S., Heger, T., Theusner, K., Jurgeit, A., Lamparter, D., Balistreri, G., Schelhaas, M., De Haan, C. a. M., Marjomäki, V., Hyypiä, T., Rottier, P. J. M., Sodeik, B., Marsh, M., Gruenberg, J., Amara, A., Greber, U., Helenius, A., and Pelkmans, L. (2012). Single-cell analysis of population context advances RNAi screening at multiple levels. *Molecular Systems Biology*, 8(579):579.

Soldatova, L. N. and King, R. D. (2005). Are the current ontologies in biology good ontologies? *Nature Biotechnology*, 23(9):1095–1098.

Sommer, C. and Gerlich, D. W. (2013). Machine learning in cell biology - teaching computers to recognize phenotypes. *Journal of Cell Science*, 126(24):5529–5539.

Sönnichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., Alleaume, A.-M., Artelt, J., Bettencourt, P., Cassin, E., Hewitson, M., Holz, C., Khan, M., Lazik, S., Martin, C., Nitzsche, B., Ruer, M., Stamford, J., Winzi, M., Heinkel, R., Röder, M., Finell, J., Häntsch, H., Jones, S. J. M., Jones, M., Piano, F., Gunsalus, K. C., Oegema, K., Gönczy, P., Coulson, A., Hyman, a. a., and Echeverri, C. J. (2005). Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature*, 434(7032):462–469.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Research*, 34(Database issue):D535–D539.

Sternberg, S. H. and Doudna, J. A. (2015). Expanding the Biologist's Toolkit with CRISPR-Cas9. *Molecular Cell*, 58(4):568–574.

Tirmizi, S., Aitken, S., Moreira, D. A., Mungall, C., Sequeda, J., Shah, N. H., and Miranker, D. P. (2011). Mapping between the OBO and OWL ontology languages. *Journal of Biomedical Semantics*, 2(Suppl 1):S3.

Trabuco, L. G., Betts, M. J., and Russell, R. B. (2012). Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. *Methods*, 58(4):343–348.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, pages 1–32.

Wang, X., Castro, M. A., Mulder, K. W., and Markowetz, F. (2012). Posterior association networks and functional modules inferred from rich phenotypes of gene perturbations. *PLOS Computational Biology*, 8(6):1–16.

Westermarck, J., Ivaska, J., and Corthals, G. L. (2013). Identification of Protein Interactions Involved in Cellular Signaling. *Molecular & Cellular Proteomics*, 12(7):1752–1763.

Wild, F. (2015). *lsa: Latent Semantic Analysis*. R package version 0.73.1.

Williams, E., Moore, J., Li, S. W., Rustici, G., Tarkowska, A., Chessel, A., Leo, S., Antal, B., Ferguson, R. K., Sarkans, U., Brazma, A., Carazo Salas, R. E., and Swedlow, J. R. (2017). Image Data Resource: a bioimage data integration and publication platform. *Nature Methods*, (November 2016).

Wilson, C. J. (2005). Identification of a Small Molecule That Induces Mitotic Arrest Using a Simplified High-Content Screening Assay and Data Analysis Method. *Journal of Biomolecular Screening*, 11(1):21–28.

Xiang, Z., Mungall, C., Ruttenberg, A., and He, Y. (2011). Ontobee: A linked data server and browser for ontology terms. *CEUR Workshop Proceedings*, 833:279–281.

Xu, T., Du, L., and Zhou, Y. (2008). Evaluation of go-based functional similarity measures using s. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 9(1):472.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabási, A.-l., Tavernier, J., Hill, D. E., and Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*, 322(5898):104–110.

Zhong, Q., Busetto, A. G., Fededa, J. P., Buhmann, J. M., and Gerlich, D. W. (2012). Unsupervised modeling of cell morphology dynamics for time-lapse microscopy. *Nature Methods*, 9(7):711–713.

Zhou, T., Ren, J., Medo, M., and Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Phys Rev E Stat Nonlin Soft Matter Phys.*, 76(4 Pt 2):046115.

Zimek, A., Schubert, E., and Kriegel, H.-p. (2012). A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Statistical Analysis and Data Mining*, 5(5):363–387.

Anexo 1

Serrano-Solano *et al. BMC Bioinformatics* (2017) 18:96
DOI 10.1186/s12859-017-1503-5

BMC Bioinformatics

RESEARCH ARTICLE Open Access

How can functional annotations be derived from profiles of phenotypic annotations?

Beatriz Serrano-Solano[1], Antonio Díaz Ramos[2], Jean-Karim Hériché[3] and Juan A. G. Ranea[1,4*]

Abstract

Background: Loss-of-function phenotypes are widely used to infer gene function using the principle that similar phenotypes are indicative of similar functions. However, converting phenotypic to functional annotations requires careful interpretation of phenotypic descriptions and assessment of phenotypic similarity. Understanding how functions and phenotypes are linked will be crucial for the development of methods for the automatic conversion of gene loss-of-function phenotypes to gene functional annotations.

Results: We explored the relation between cellular phenotypes from RNAi-based screens in human cells and gene annotations of cellular functions as provided by the Gene Ontology (GO). Comparing different similarity measures, we found that information content-based measures of phenotypic similarity were the best at capturing gene functional similarity. However, phenotypic similarities did not map to the Gene Ontology organization of gene function but to functions defined as groups of GO terms with shared gene annotations.

Conclusions: Our observations have implications for the use and interpretation of phenotypic similarities as a proxy for gene functions both in RNAi screen data analysis and curation and in the prediction of disease genes.

Keywords: Ontology, Cellular phenotype, Biological network, Cluster analysis

Background

A central tenet of experimental approaches to assigning functions to genes posits that genes involved in the same biological process show similar loss-of-function phenotypes. This provides the rationale for performing loss-of-function genetic screens and is used by the Gene Ontology consortium in their gene annotation process (i.e. for annotations with evidence code IMP: Inferred from Mutant Phenotype, The Gene Ontology Evidence Tree[1]). Systems microscopy approaches, defined as the combination of recent developments in microscopy automation with automated image analysis and data mining [1], now allow for systematic exploration of the gene loss-of-function phenotypic space and large scale RNAi screens have given us phenotypic information for thousands of genes (e.g. [2–4]). In contrast to more traditional experiments that have been addressing a single phenotype closely associated with a function, systems microscopy approaches increasingly use phenotypic profiling, the description of phenotypes by multi-parameter measurements. While this increases the amount of usable information, the cost is that functional associations become less evident. The process of converting phenotypic annotations to functional annotations therefore remains a manual one, due to the free-text nature of many phenotypic descriptions and to the difficulty of assessing phenotypic similarity (i.e. how similar should two phenotypes be in order to infer the same function?) in particular across different experiments. As a consequence, large RNAi screens performed in human cells haven't been used to annotate genes with Gene Ontology terms from the biological process domain and this contributes to a lower level of experimentally-supported annotations of genes cellular functions than the number of reported functional assays.

*Correspondence: ranea@uma.es
[1]Department of Molecular Biology and Biochemistry, University of Málaga, Boulevard Louis Pasteur, 29071 Málaga, Spain
[4]CIBER de Enfermedades raras (CIBERER), Madrid, Spain
Full list of author information is available at the end of the article

In this context, automating the identification of similar cellular phenotypes and their assignment to relevant cellular processes across experiments would increase the level of experimentally-supported functional annotations. The recently developed Cellular Phenotype Ontology (CPO, [5]) and Cellular Microscopy Phenotype Ontology (CMPO, [6]) attempt to fill a gap in the domain coverage of existing ontologies by organizing the cellular phenotype domain into a consistent ontological structure. Replacing free-text phenotypic description in RNAi screens with well-defined ontology terms makes automatic evaluation of phenotypic similarity possible. However, to automatically convert phenotypic annotations, in particular phenotypic profiles, to functional annotations, we need to understand how phenotypes and functions are related. As most screens report hit lists usually found enriched in genes involved in relevant biological processes using GO annotations, we can expect phenotypic similarity to correlate with or be indicative of participation in similar cellular processes. We wondered how phenotypic profiles generated by combining annotations from multiple screens could be exploited to automate and/or refine gene functional annotations. Note that our goal is not to remine the screens to infer gene function but rather to explore whether and how the phenotypic annotations resulting from these screens can be related to Gene Ontology biological process terms.

Methods

Gene phenotypes

Gene loss-of-function phenotypes were obtained from the following large siRNA-based gene silencing experiments performed in human cells: Mitocheck [2], EMBL secretion screen [4], EMBL chromosome condensation screen [7], Copenhagen DNA damage screen [8], CellMorph [3]; and RNAi screens GR00290-A (regulation of centriole biogenesis, [9] and GR00053-A (genome stabilization by phosphorylation of the histone H2AX, [10] from the GenomeRNAi database [11]. It is noteworthy that none of these screens have been used for making biological process annotations in GO (as evidenced by the fact that none of the corresponding papers are cited as source of annotation) despite the data having been available for several years. The cellular functions covered by these screens are diverse and include cell proliferation, cell death, cell motility, mitosis, protein secretion, DNA damage and centriole formation. However, this list is not exhaustive as some screens (e.g. CellMorph, MitoCheck) report phenotypes not obviously associated with a declared target biological function.

The compilation of all data from these separate experiments, gives a set of 36 unique cellular phenotypes (Table 1) associated with 4198 Entrez genes (see Additional file 1: Table S1). Most genes have been tested in more than one screen, and the screens include non-overlapping sets of phenotypes. As our goal is to explore how phenotypic annotations are linked to GO cellular process terms, we used phenotypic annotations resulting from the screens as made available in the corresponding papers. Relationships between genes and phenotypes from different assays were integrated into a binary matrix recording the presence (value 1) or absence (value 0) of a phenotype for each gene (Table 2). Note that 0 is also used where genes have not been tested in a screen. To assess whether this affected our results, we tested the effect of sparsity by replacing a proportion (5, 10, 20 and 30%) of randomly selected 1 s with 0 s. A visual overview of the data matrix is presented in Additional file 2: Figure S1.

Ontologies and annotations

We used two formal ontologies to perform our study: the Gene Ontology (GO) [12] and the Cellular Microscopy Phenotype Ontology (CMPO)[2]. We selected for our study the GO branch of cellular process (root term GO:0009987), which is the ontological domain closer to the cellular phenotypes captured in the screens. The terms hierarchy was extracted from the OBO file released on 2015-09-26. Gene Ontology annotations of genes were downloaded from the GO web site[3] (see Additional file 3: Table S2) and extracted from the file with validation date: 09/16/2015, removing electronically-inferred annotations (IEA). To ensure that the genes with phenotype did not form a biased set of GO annotations, we verified that the distribution of information content of the terms used to annotate the genes with phenotypes was the same as for all annotated genes (Fig. 1).

CMPO gene annotations were retrieved from the cellular phenotype database[4] [13](see Additional file 4: Table S3). Compared to a vocabulary of phenotypes, the ontology has the advantage of formalizing the relationships between the phenotypes. For example, the ontology allows to infer that the phenotypes "chromosome segregation defect" and "metaphase arrest" are both mitotic phenotypes.

Similarity measures

Similarity measures used in this study are shown in Table 3. Euclidean and correlation distances were computed using the R core package *stats*, for cosine we used *lsa* [14] and for Jaccard *prabclus* [15]. Hamming, Cohen's kappa and TF-IDF [16] were also coded in R. For dimensionality reduction, we applied the *logisticPCA()* function of the R package *logisticPCA* to extract 10 principal components and correlation, cosine and Euclidean similarities were computed in this new space. To take advantage of the phenotype ontology, we also computed several measures of semantic similarity using the R package *dnet* [17].

Table 1 Set of 36 phenotypes obtained from the listed siRNA experiments sorted by its CMPO identifier

Experiment	Description	Phenotypes	IDs in CMPO
CellMorph [3]	Genome-wide RNAi screen that examines changes in the morphology of individual HeLa cells within cell populations.	- Decreased cell number	CMPO:0000052
		- Cell with projections	CMPO:0000071
		- Elongated cell	CMPO:0000077
		- More lamellipodia cells	CMPO:0000083
		- Increased number of actin filament	CMPO:0000105
		- Round cell	CMPO:0000118
		- Increased cell size	CMPO:0000128
		- Decreased cell size	CMPO:0000129
		- Bright nuclei	CMPO:0000154
		- Metaphase arrested	CMPO:0000305
		- Increased cell size in population	CMPO:0000340
MitoCheck [2]	Genome-wide RNAi screen for genes required for chromosome segregation in HeLa cells. The screen also reports genes involved in other processes such as cell movement.	- Cell death	CMPO:0000030
		- Increased nucleus size	CMPO:0000140
		- Graped micronucleus	CMPO:0000156
		- Abnormal nucleus shape	CMPO:0000157
		- Mitosis delayed	CMPO:0000202
		- Binuclear cell	CMPO:0000213
		- Absence of mitotic chromosome decondensation	CMPO:0000216
		- Increased cell movement speed	CMPO:0000236
		- Increased cell movement distance	CMPO:0000237
		- Proliferating cells	CMPO:0000241
		- Metaphase delayed	CMPO:0000307
		- Abnormal chromosome segregation	CMPO:0000326
		- Prometaphase delayed	CMPO:0000344
		- Increased variability of nuclear shape in population	CMPO:0000345
		- Mitotic metaphase plate congression	CMPO:0000348
EMBL secretion [4]	Genome-wide RNAi screen for interference with ER-to-plasma membrane transport of the secretory cargo protein tsO45G in HeLa cells.	- Increased rate of protein secretion	CMPO:0000246
		- Mild decrease in rate of protein secretion	CMPO:0000318
		- Strong decrease in rate of protein secretion	CMPO:0000319
		- Decreased rate of intracellular protein transport	CMPO:0000346
GR00053 [10]	Genome-wide RNAi screen for genes involved in DNA damage responses in HeLa cells.	- Increased number of site of double-strand break	CMPO:0000182
GR00290 [9]	Genome-wide RNAi screen for genes regulating centriole formation in HeLa cells.	- Increased centriole replication	CMPO:0000361
		- Decreased centriole replication	CMPO:0000362
Copenhagen DNA damage Ubiquitin [8]	RNAi screen of >1300 genes involved in the ubiquitin-proteasome system or encoding zinc-finger proteins looking for modulators of cellular responses to ionizing radiation in HeLa and U2OS cells.	- Decreased number of site of double-strand break	CMPO:0000181
EMBL chromosome condensation [7]	RNAi screen of 100 bioinformatically-selected genes for changes in mitotic prophase duration in HeLa cells.	- Increased duration of mitotic prophase	CMPO:0000328
		- Decreased duration of mitotic prophase	CMPO:0000329

Table 2 Binary matrix for gene-phenotype association

Gene	Decreased cell number (CMPO:0000052)	Cell with projections (CMPO:0000071)	…	Mitotic metaphase plate congression (CMPO:0000348)
57147 (SCYL3)	1	0	…	0
2268 (FGR)	1	0	…	1
22875 (ENPP4)	0	1	…	0
…	…	…	…	…
5439 (POLR2J)	1	0	…	1

Presence and absence of a phenotype after inhibition of each gene is represented by values 1 and 0, respectively

Comparison of phenotypic similarity measures

To evaluate how the similarity measures related to each other, similarities between pairs of genes were computed for each measure. Pearson's correlation coefficient (PCC) between the measures was then computed from these sets of values. Hierarchical clustering was performed by average linkage using the *hclust* R package with 1-PCC as distance measure.

To rank the similarity measures in relation to their ability to capture gene function we used protein interaction as a proxy for functional relationships between genes. To this end, we first ranked the similarity measures by their ability to distinguish between interacting and non interacting gene pairs using the area under the ROC curve (AUC). In this context, the AUC can be interpreted as the probability that the similarity measure ranks an interacting gene pair higher than a non-interacting one. As positive interacting pairs, we used physical protein interactions from Intact [18], MIPS [19], DIP [20] and BIOGRID [21] that have been reported by two different experimental methods and curated interactions from Reactome [22, 23]. As negative interactions, we used the curated negative interactions from the MIPS Negatome [24] and Trabuco et al. [25]. The AUCs were computed using the R package *pROC* [26].

We also computed a score for each measure as the number of genes whose most phenotypically similar gene is also an interaction partner in the iRef index protein interaction data (release 14.0, April 7th, 2015) [27]. For each measure, the nearest neighbor of each gene was identified (ties were broken at random) and the measure's score was incremented by one if the two genes formed a known interacting pair in the iRef index. To assess the statistical significance of the score, the probability of having the same or better score from a random selection of protein interactions was computed from the hypergeometric distribution using the *phyper()* function in R as follows: We considered 4198 genes making a total of $4198 * (4198 - 1)/2$ possible interactions of which 29649 were present in iRef index. For a given measure of similarity, we tested 4198 interactions (one for each gene). Therefore, the probability of having a score of x or better by selecting the interactions randomly is given by $1 - phyper(x - 1, 29649, 4198 * (4198 - 1)/2 - 29649, 4198)$.

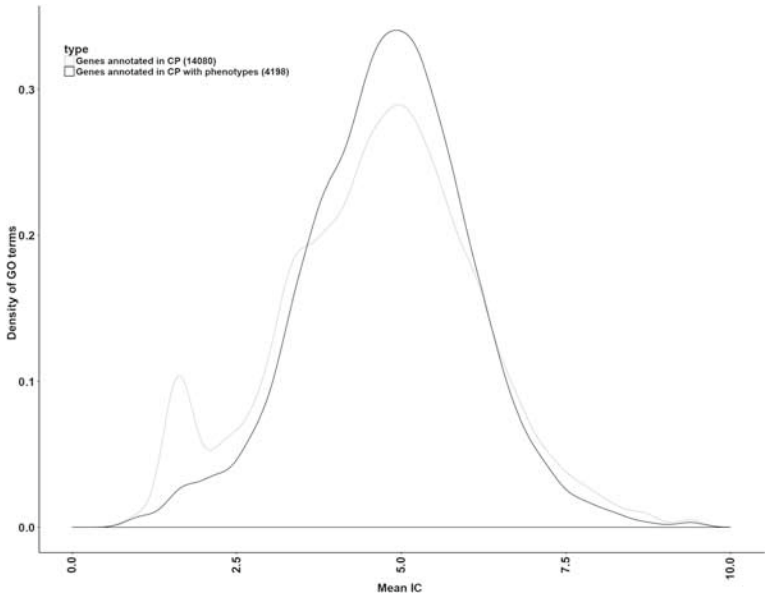


Fig. 1 Distribution of information content (IC) of the terms annotating genes with phenotypes (*black*) and all the terms in cellular process (*grey*). For each level of specificity represented by the information content (IC), the *curves* represent the proportion of genes annotated with terms of this level in all the annotated genes versus the subset of genes with phenotypes

Table 3 Similarity measures used in this study

Name	Formula
Euclidean similarity	$s^2(g_1,g_2) = \frac{1}{1+(x_{g1}-x_{g2})(x_{g1}-x_{g2})'}$
Correlation similarity	$s(g_1,g_2) = \frac{(x_{g1}-\bar{x}_{g1})(x_{g2}-\bar{x}_{g2})'}{\sqrt{(x_{g1}-\bar{x}_{g1})(x_{g1}-\bar{x}_{g1})'}\sqrt{(x_{g2}-\bar{x}_{g2})(x_{g2}-\bar{x}_{g2})'}}$ where $\bar{x}_{g1} = \frac{1}{n}\sum_{p\in P} x^p_{g1}$ and $\bar{x}_{g2} = \frac{1}{n}\sum_{p\in P} x^p_{g2}$
Cosine similarity	$s(g_1,g_2) = \frac{x_{g1}x'_{g2}}{\sqrt{x'_{g1}x_{g1}}\sqrt{x'_{g2}x_{g2}}}$
Hamming similarity	$s(g_1,g_2) = \frac{x^p_{g1}=x^p_{g2}}{n}$
Jaccard similarity	$s(g_1,g_2) = 1 - \frac{\left[\left(x^p_{g1}\neq x^p_{g2}\right)\wedge\left(\left(x^p_{g1}\neq 0\right)\vee\left(x^p_{g2}\neq 0\right)\right)\right]}{\left(x^p_{g1}\neq 0\right)\vee\left(x^p_{g2}\neq 0\right)}$
Cohen's kappa	$s(g_1,g_2) = \frac{p_0-p_c}{1-p_c}$ where: - p_0 is the proportion of terms common to profiles g_1 and g_2, and - p_c is the proportion of terms common to profiles g_1 and g_2 expected by chance.
TF-IDF similarity	$s(g_1,g_2) = \max_{p\in P}\left\{x^p_{g1}x^p_{g2}IDF(p)\right\}$ where $IDF(p) = log\frac{n_G}{1+\sum_{g\in G}x^p_g}$
Resnik's semantic similarity	$s(t_1,t_2) = IC(t_{MICA})$ where: - the Most Informative Common Ancestor is $t_{MICA} = argmax_{t\in S(t_1,t_2)}IC(t)$, - the information content (IC) of a term t is $IC(t) = -log(p(t))$, - the probability of a term t is $p(t) = \frac{annotations(t)}{totalAnnotations}$, and - $S(t_1,t_2)$ is the set of common ancestors of t_1 and t_2.
Lin's semantic similarity	$s(t_1,t_2) = \frac{2\cdot IC(t_{MICA})}{IC(t_1)+IC(t_2)}$
Schlicker's semantic similarity	$s(t_1,t_2) = \frac{2\cdot IC(t_{MICA})}{IC(t_1)+IC(t_2)}\cdot(1-p(t_{MICA}))$
Jiang's semantic similarity	$s(t_1,t_2) = 1 + 2\cdot IC(t_{MICA})\,(IC(t_1)+IC(t_2))$
Pesquita's semantic similarity	$s(t_1,t_2) = \frac{\sum_{t\in S(t_1,t_2)} IC(t)}{\sum_{t\in P(t_1,t_2)} IC(t)}$ where: - $P(t_1,t_2)$ is the set of ancestors of either t_1 or t_2.

G is the full set of genes ($n_G = 4198$) and P is the set of 36 (n_P) phenotypes. x_g denotes the phenotypic profile of gene g with $x^p_g = 1$ if g shows phenotype p, $x^p_g = 0$ otherwise

Annotation-driven approach

Following the approach by Glass and Girvan [28], a bipartite graph was constructed, for functions and phenotypes respectively, by setting an edge between two GO terms (resp. CMPO terms) if they shared at least a gene and the edge was weighted by the number of genes shared. Because high level terms inherit genes from their child terms, term degrees are biased. To compensate for this, we normalized edge weights by the union of the genes belonging to the two terms. We then grouped terms by spectral clustering using the normalized cut objective function [29] with an arbitrary number of clusters, set to 13 for CMPO and 140 for GO. GO terms clusters were obtained by first partinioning the graph into 100 clusters then partitioning again the two largest clusters into 33 and 9 clusters. As noted by Glass and Girvan [28], different numbers of clusters correspond to different levels of specificity. We chose the number of GO terms clusters so that most clusters would be linked to phenotypes. The number of CMPO terms clusters was chosen to produce a reasonable distribution of cluster sizes minimizing the number of clusters with only one single term. Increasing the number of clusters leads to an increase in the number of clusters containing only one term.

Correction for multiple testing

P-values were corrected for multiple testing using the R function *p.adjust()* with the Benjamini and Hochberg method.

Results

Comparison of phenotypic similarity measures

As we wished to link phenotypic similarity to gene function, the first question we addressed is which measure of phenotypic similarity to use for the task. Similarity between phenotypic profiles has typically been assessed using feature vector-based similarity measures such as correlation [30, 31] or cosine (e.g. [32, 33]). Due to their binary nature, profiles can also be compared using character-based (binary) similarity measures. For example, the main component of the PhenoBlast algorithm for retrieving profiles similar to a query [34] is the number of matches in the binary string. PhenoBlast also recognizes that some phenotypes may be more informative than others and one of its components is the probability of observing a given combination of shared phenotypes by chance. Combining these two components into one measure leads to Cohen's kappa measure of similarity between two profiles. The intuition that some phenotypes are more informative than others can be formalized by using information content-based similarity measures. Here, information content refers to the specificity of a phenotype. Typically, a phenotype is considered more specific if it is less often observed e.g. cell death, a widely observed phenotype, is considered less specific than mitotic delay which is more rarely observed. This leads to TF-IDF similarity measures in which phenotypes are weighted by the inverse of their frequency of occurrence in the data [35]. The availability of CMPO now also allows for a semantic information content-based approach to phenotypic similarity, analoguous to what has been used with Gene

Ontology annotations (e.g. Resnik's similarity measure). In an ontology, the information content of a term also takes into account the structure of the ontology such that child terms are more specific than their parents. When working with feature vectors of high dimension, it is sometimes beneficial to compute vector-based similarity measures in a reduced dimensional space. As phenotypic profiles are high dimensional vectors, we also wondered if a dimensionality reduction approach would be beneficial and applied logistic PCA, an extension of standard PCA to binary data, to compute vector-based phenotypic similarities in a reduced dimensional space.

Given these different ways in which to measure phenotypic similarities, we wondered whether they were equivalent in ranking genes based on their phenotype profiles. To answer this question, we computed the correlation coefficient between phenotypic similarities obtained with the different measures and performed hierarchical clustering. The resulting dendrogram (Fig. 2) shows that the similarity measures fall into two main groups with the information content-based semantic similarity measures (Resnik, Schlicker, Lin, Jiang and Pesquita) distinctly separated from the feature vector-based measures (cosine, Euclidean, correlation, Jaccard and Hamming), with Cohen's kappa occupying an intermediate position, confirming our intuition that these groups of measures assess phenotypic similarity in different ways. We next asked whether this difference was meaningful with respect to biological function. To test this, we used protein interactions as a proxy for biological function, i.e. two interacting proteins are taken as indication that the corresponding genes are involved in the same function [36]. This means that, for a relevant measure, phenotypically similar genes are expected to be enriched in protein interactions. We tested this in two ways. First, we assessed the ability of each measure to distinguish between interacting and non-interacting gene pairs by computing the area under the ROC curve (AUC) using high-confidence physical protein-protein interactions as positive set and curated non-interacting protein pairs as negative set. In this context, the AUC is the probability that the similarity measure ranks an interacting gene pair higher than a non-interacting one. A similarity measure with no discriminating power has an AUC of 0.5 and higher values indicates increasingly better discriminative power. Using this approach, the best similarity measures are Resnik's and Schlicker's with the other semantic similarity measures outperforming the character- and vector-based measures (Table 4). Therefore, using semantic similarity measures, phenotypic profiles of interacting genes are overall more similar than for non-interacting gene pairs.

In a second approach, for each similarity measure, we identified the nearest (i.e. most similar) neighbour of each gene and tested whether the two genes were known interaction partners. To compare the phenotypic similarity measures, we then ranked them by the number of interactions retrieved in this way (Table 4). With this approach, TF-IDF and Resnik's similarity performed best. Other semantic similarity measures and most feature-based measures (Euclidean, Jaccard and cosine) were not better than a random selection of protein interactions indicating that these phenotypic similarity measures may not adequately capture functional relationships. Dimensionality reduction as obtained by logistic PCA did not

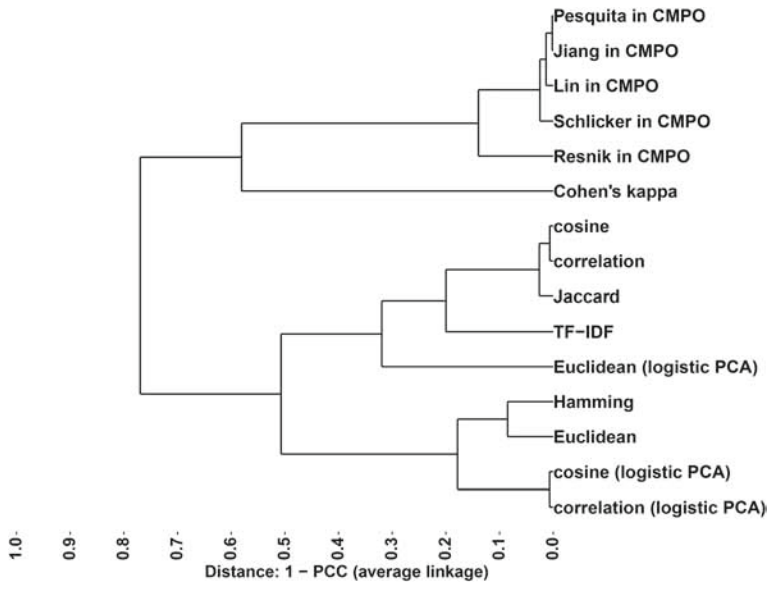


Fig. 2 Hierarchical clustering of phenotypic similarity measures based on Pearson correlation distance

Table 4 Similarity measures sorted by area under the ROC curve (AUC)

Measure	AUC	Protein interactions	*p*-value
Resnik in CMPO	0.56	24	0.0102
Schlicker in CMPO	0.56	12	0.7512
Lin in CMPO	0.55	11	0.8332
Cohen's kappa	0.54	27	0.0015
Pesquita in CMPO	0.54	14	0.5494
Jiang in CMPO	0.54	11	0.8332
TF-IDF	0.53	25	0.0055
Euclidean	0.53	16	0.3433
correlation	0.52	22	0.0311
Hamming	0.52	21	0.0513
cosine	0.49	13	0.6545
Jaccard	0.49	13	0.6545
Euclidean (logistic PCA)	0.46	25	0.0055
correlation (logistic PCA)	0.45	19	0.1242
Cosine (logistic PCA)	0.45	14	0.5494

The second column represents the number of nearest neighbour gene pairs who are also protein interaction partners, and the third one, the *p*-values (computed from the hypergeometric distribution) that the number of observed interacting pairs is due to chance

improve performance of the vector-based measures indicating that linear combinations of phenotypes are unlikely to capture links to function. Therefore, across the two tests, Resnik's similarity measure appears the most consistent at associating similar phenotypes with interacting proteins. Other semantic similarity measures may have been negatively influenced by the sparsity of the CMPO ontology due to their attempts at accounting for more of the ontology structure than Resnik's measure. For example, Lin's and Jiang's measures are particularly sensitive to variations in the ontology structure because they take into account the density and the level of the terms whereas Resnik's measure only considers the lowest common ancestor and is thus comparatively more robust.

Relationship between GO cellular process annotations and phenotypes

Having identified a suitable measure of phenotypic similarity, we set to explore how gene functions relate to phenotypes more directly. If phenotypes are predictive of biological functions, we expect that pairs of genes with similar phenotypes will have similar functions. Since gene functions have been standardized using the Gene Ontology, gene functional similarity was computed using Resnik's semantic similarity between GO terms, a measure generally found to be the best for this purpose [37]. To assess links between gene phenotypic similarity and gene semantic similarity in GO, we plotted GO semantic similarities versus CMPO semantic similarities for the RNAi screen data (Fig. 3a), excluding genes with no functional annotation in GO. The distribution of functional similarity values is the same for all levels of phenotypic similarity except the highest, which showed a trend towards higher functional similarity. Although weak, this effect is robust as it is still observed when removing up to 30% of the phenotypic annotations (see Additional file 5: Figure S2) and does not appear to be due to chance because random assignment of GO similarity values to high-scoring CMPO gene pairs resulted in a lower average GO similarity (Additional file 6: Figure 3). While this matched our expectation that specific phenotypes are associated with specific functions, this represented only a small fraction of the genes (20/4198) and for most genes, phenoypes do not appear to be good indicators of function.

One possible explanation for this result is that several functions could share the same phenotype. If that were the case, then we would predict that similar functions would still lead to similar phenotypes. We would then expect that two genes involved in the same cellular process would have similar phenotypes. However, this is not the case as genes with high functional similarity are not more likely to have high phenotypic similarity (Fig. 3b). This lack of correlation between function and phenotype was also observed for the other phenotypic similarity measures tested, indicating that this was not an effect of the phenotypic similarity measure used. This effect is also observed when electronically-inferred annotations are included (see Additional file 6: Figure S3).

So neither considering the most informative phenotypic term nor the whole phenotypic profile gives any indication of function. This result is counter-intuitive since the premise of most screens is that genes with the same biological function would give the same loss-of-function phenotype or phenotypic profile. We hypothesized that perhaps even in screens which relied on profiling, each phenotype is individually indicative of a function. To test whether this also holds across screens, we averaged the semantic similarity in GO for all pairs of genes showing a particular phenotype. Then, we compared this average to that obtained from 100 datasets generated by randomly shuffling the associations between genes and phenotypes while keeping the number of links per phenotype unchanged. A total of 8 out of 36 (25%) phenotypes gave a statistically significant signal (FDR-corrected p-value$\leqslant$0.01) for having their actual functional similarity between genes above that obtained by randomization (Fig. 4). Half of these significant phenotypes correspond to CMPO terms with high information content indicating that only specific phenotypes tend to associate with highly similar GO functional annotations. While these results conform to our intuition, the only practical rule that can

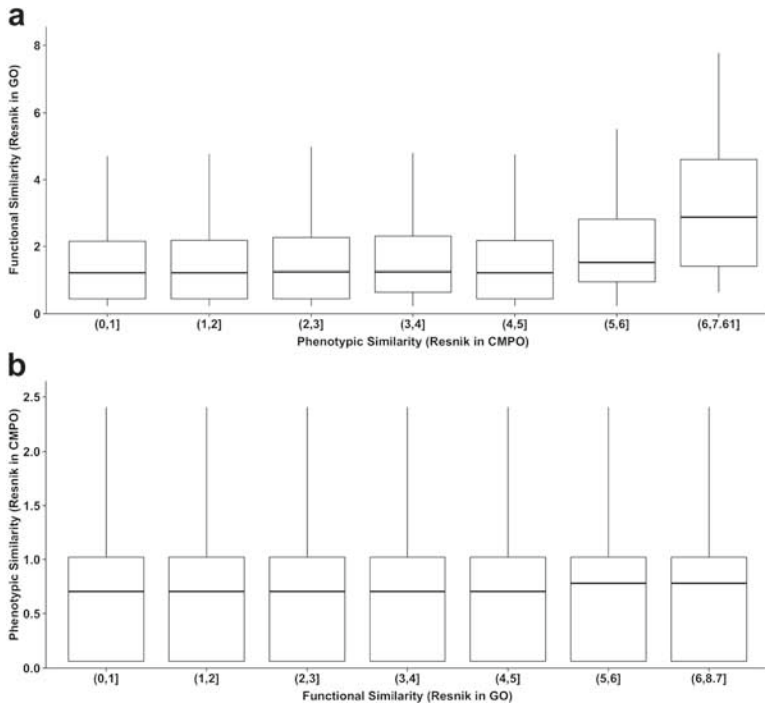


Fig. 3 Distributions of functional and phenotypic similarities. The *box* represents the *upper* and *lower* quartiles and the median is represented by the *black* line inside the *box*. **a** Phenotypic similarity in CMPO versus functional similarity in GO. **b** Functional similarity in GO versus phenotypic similarity in CMPO

be derived for automatically converting phenotypic annotations to functional annotations is that only phenotypes with CMPO semantic similarity over some threshold are indicative of similar cellular function.

Gene annotation-driven phenotypic and functional similarity

The above results suggested that the Gene Ontology structure does not adequately capture the functional relationships that underlie phenotypic similarity. An alternative way of organizing GO terms has recently been proposed by Glass and Girvan [28]. In this scheme, a term graph is generated by linking terms based on the genes annotated with them. Thus, two terms are more similar the more genes they share (i.e. the more genes are annotated with both terms). Biological functions can then be defined as groups of similar terms by applying a clustering algorithm to the term graph (Fig. 5).

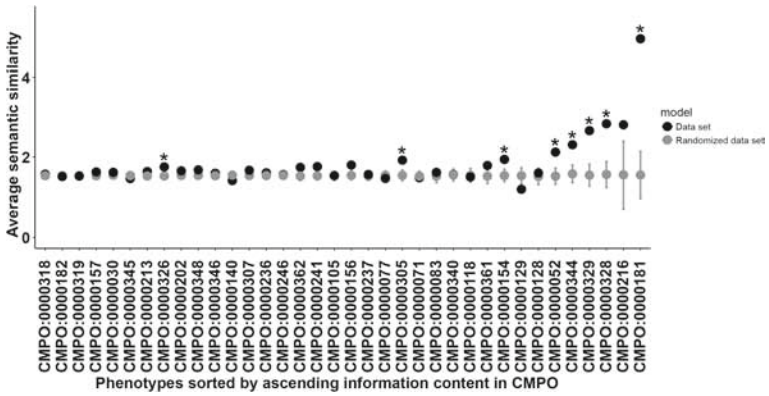


Fig. 4 Average semantic similarity in GO between genes sharing a particular phenotype (*black*). Randomization of the relationships between phenotypes and genes represents the null model (*grey*). Phenotypes with genes having high functional similarity (FDR-corrected *p*-values$\leqslant$0.01) are marked with *. Phenotypes are sorted on the *X* axis by ascending information content in CMPO. CMPO descriptions for the identifiers are in Table 1

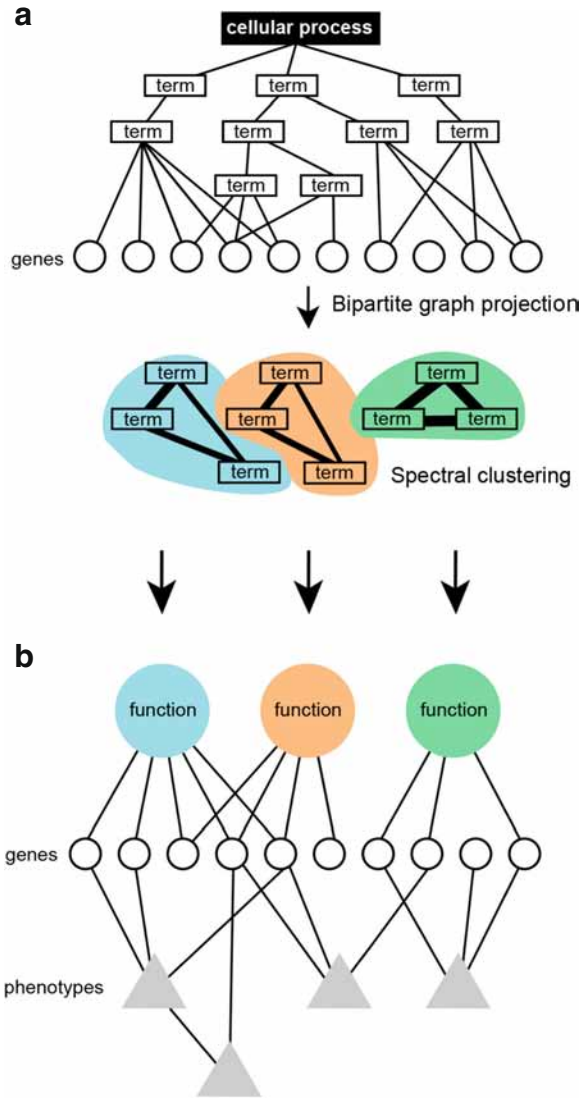


Fig. 5 Annotation-driven cellular function definition. **a** Genes (*circles*) are annotated with cellular process ontology terms (*rectangles*). After bipartite graph projection, links between terms are weighted according to the number of genes shared (*line width*). Then, terms are grouped using spectral clustering. **b** Clusters of functional terms (*coloured circles*) are linked to phenotypes (*triangles*) by shared genes

In this scheme, a function can be seen as being represented by a signature of co-occurring terms. We wondered if this approach would allow us to recover a broader relationships between functions and phenotypes. To test this, we grouped the cellular process GO terms into 140 clusters. To assess whether this new definition of function captured phenotypic similarity we computed Resnik's similarity between CMPO terms associated with genes within each cluster (Fig. 6a). Excluding functional clusters not linked to phenotypes, 77%(45/58) of functional clusters had high phenotypic similarity that could not be explained by chance assignment of GO terms to clusters (FDR-corrected p-value$\leqslant$ 0.01, Fig. 6a). Therefore cellular functions derived from shared gene annotations were associated with phenotypic similarity. To test whether similar phenotypes reflected similar functions, we defined a phenotypic terms graph in the same way and grouped the phenotypes into 13 clusters. Each of these phenotypic cluster can be viewed as a phenotype characterized by a signature of co-occurring phenotypic descriptors. As above, for each phenotypic cluster, we computed Resnik's similarity between GO terms within clusters. Again, except for clusters with no GO annotations, we observed that functional similarity was higher in phenotypic clusters than can be explained by random phenotype assignments to clusters (Fig. 6b). This indicated that this definition of phenotype was able to recover functional similarity in GO. Therefore, functions defined by groups of GO terms sharing associated genes tend to map to CMPO terms better than functions defined by individual GO terms and conversely, phenotypes defined by groups of CMPO terms sharing associated genes map better to GO terms than phenotypes defined by individual CMPO terms. While the details of how phenotypes and functions are defined is subject to changes in both CMPO and GO, the strong association between phenotypes and functions is robust as it only depends on the annotated genes.

Discussion

The large amount of cellular phenotypic annotations coming from high-throughput genetic screens represents a largely untapped source of information on gene function. Our aim was to understand how these phenotypes are related to gene function in the hope that principles could be derived for use in automatically converting published phenotypic annotations to functional annotations. Here, we used published cellular phenotypes from large scale RNAi screens in human cells that have been annotated with CMPO terms to explore how cellular phenotypes related to GO cellular functions. The first question we addressed was how to adequately measure phenotypic similarity such that phenotypic similarity would be correlated with functional similarity. We found that, in contrast to feature-based similarity measures, information content-based phenotypic similarity measures like Resnik's semantic similarity were best at associating high phenotypic similarity with protein interactions, suggesting that these phenotypic similarity measures were the most likely to capture functional relationships. The poor performance of character- or vector-based measures of phenotypic similarity lies at least in part in the fact that they can be misled by genes involved in the same function but having been assigned different phenotypic descriptions as for example positive and negative

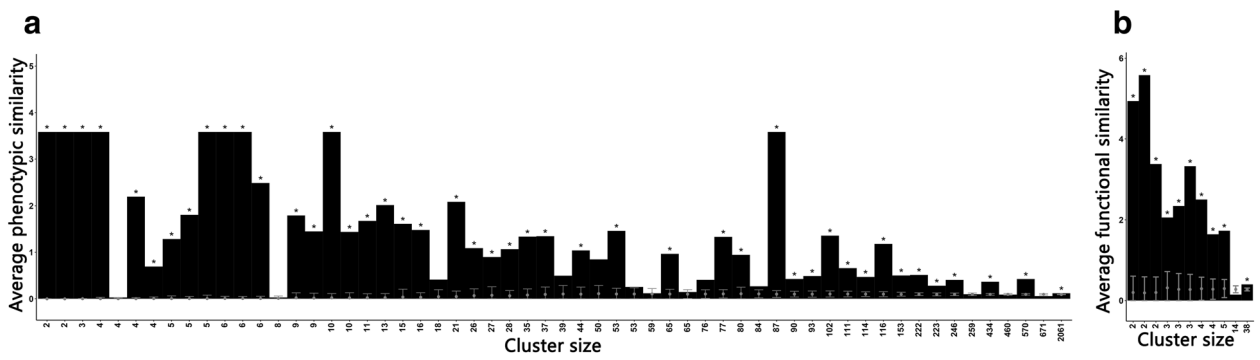


Fig. 6 Average semantic similarity between terms in clusters. Randomization of the assignments of terms in clusters are represented in *grey*. Clusters are sorted by size (i.e. number of terms). **a** Average phenotypic similarity in clusters of GO terms. **b** Average functional similarity in clusters of CMPO terms

regulators having opposite effects on a particular cellular feature. These measures are also affected by differences in phenotypic annotations of any given genes across screens as, for example, they treat 'metaphase delayed' and 'mitosis delayed' as unrelated phenotypes of the same gene. Ontology-based semantic similarity measures on the other hand do not have this problem. Measures accounting for chance occurrence of a phenotype such as TF-IDF also perform better than the character-based methods and this could be attributed to the relationship between frequency of a phenotype and its specificity, i.e. more specific phenotypes tend to be less represented in the data. However, despite semantic similarity measures looking promising, only phenotypes with high semantic similarity in CMPO were associated with high functional similarity of GO cellular function annotations. To use this observation for automatically converting phenotypes into GO functional annotations, one would need to define a threshold of CMPO semantic similarity above which function assignment becomes reliable but how to select this threshold is unclear because it is liable to change when the ontology is expanded. Another downside is that only a small fraction of genes with phenotypes could be annotated with cellular functions in this way.

We therefore wondered if another approach could make better use of the information. Defining cellular functions as groups of co-occurring GO terms allowed us to recover a stronger link between phenotypic similarity and function. Conversely, defining phenotypes as sets of co-occurring CMPO terms allowed us to link these phenotypes to similar functions in GO. Therefore, with these definitions, similar cellular functions do lead to similar phenotypes and similar phenotypes are indicative of similar functions. Our results extend the observation by Glass and Girvan [28] that cancer signatures can associate with GO term communities but not branches of the Gene Ontology. We note that, by requiring as input a list of functionally-related genes, some network-based gene prioritization algorithms such as FUN-L [38] and GeneMANIA [39] implicitly rely on this definition of biological function and in light of our findings, this may contribute to their success in enriching candidate genes in the desired phenotypes.

Our observations have several practical implications. First, they suggest that clustering of phenotypic profiles using naive profile vector-based metrics (as commonly done in the field of RNAi screening) is sub-optimal for predicting the function of genes because these types of measures have low correlation with functional similarity but correlation can be improved by taking into account information content of the phenotypes. Instead of clustering the genes, we propose that a more meaningful approach would be to cluster the phenotypes based on the genes annotated with them and look for enrichment in functional terms in these clusters. Genes associated with a cluster of CMPO terms can then be annotated with the corresponding functional GO terms. This is relevant to any gene annotation task whether through curation of existing data or analysis of an RNAi screen with multiple phenotypes.

A second implication concerns the integration of phenotypic information with other biological data. Several candidate gene selection methods rely on the combination of multiple sources of information to increase accuracy and coverage of functional association between human genes. So far phenotypic data from RNAi screens have not been used in these data integration schemes. While supervised machine learning methods could learn to make functional annotations from phenotypic ones, the outcome critically depends on the quality of the training set which in turn depends on how one links functional annotations to phenotypes. This is important for example to design a relevant kernel for kernel-based methods such as support vector machines. In this context, the design of meaningful kernels for phenotypic similarity would be an advantage. In our experience, and consistent

with results presented here, using standard metrics to compute similarity between phenotypic profiles leads to poor performance in retrieving functionally related genes. Our results suggests that better phenotype kernels could be derived by replacing individual phenotypes by clusters of CMPO terms derived from the annotation-based graph. In the same way, considering diseases as phenotypes, we suggest that functional similarity derived from the annotation-based clusters of GO terms could be more useful for predicting disease genes than semantic similarity-based functional similarity.

Finally, as not every single gene knock-down can reveal a phenotype, studies have turned to phenotyping genetic interactions using RNAi (e.g. [30, 40, 41]). Whether and how these can be integrated in the way we propose here is an area of future work.

Conclusions

In this work we explored how gene phenotypic annotations from RNAi screens in human cells are related to functional annotations in GO. After selecting a relevant measure to compare phenotypic profiles, we compared gene pairs similarities using GO and CMPO and found that phenotypic similarity generally did not correlate with functional similarity in GO. However, redefining functions as groups of co-occurring GO terms allowed us to recover a stronger link between phenotypes and functions. Our observations are particularly relevant in situations where phenotypic similarities are used as a proxy for inferring gene functions such as in RNAi screen data analysis and curation, in integrating phenotypic data with other data and in the prediction of disease genes.

Endnotes

[1] http://www.geneontology.org/GO.evidence.tree.shtml.

[2] http://www.ebi.ac.uk/cmpo.

[3] http://geneontology.org/page/download-annotations.

[4] http://www.ebi.ac.uk/fg/sym.

Additional files

Additional file 1: Table S1. List of genes. (CSV 47 kb)

Additional file 2: Figure S1. Heatmap of the Gene x Phenotype matrix. (PNG 7680 kb)

Additional file 3: Table S2. Gene annotations to GO terms. (CSV 2048 kb)

Additional file 4: Table S3. Gene annotations to CMPO terms. (CSV 180 kb)

Additional file 5: Figure S2. Distribution of functional similarity in GO versus phenotypic similarity (and vice versa) for different levels of sparsity. (PDF 250 kb)

Additional file 6: Figure S3. Distribution of average semantic similarities between genes for those pairs with high phenotypic similarity (>6) after random assignment of GO similarity values. (PDF 45 kb)

Abbreviations

AUC: Area under the ROC curve; CMPO: Cellular microscopy phenotype ontology; CPO: Cellular Phenotype Ontology; EMBL: European molecular biology laboratory; GO: Gene ontology; IC: Information content; IDF: Inverse document frequency; IEA: Inferred from electronic annotation; IMP: Inferred from mutant phenotype; OBO: Open biomedical ontologies; PCA: Principal component analysis; PCC: Pearson's correlation coefficient; RNAi: RNA interference; ROC: Receiver operating characteristic; siRNA: small interfering RNA; TF-IDF: Term frequency - inverse document frequency

Acknowledgements

The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the SCBI (Supercomputing and Bioinformatics) center of the University of Málaga. We also thank Gabriella Rustici and Rocío Rodríguez-López for technical support while this study was being conducted and James R. Perkins for his help revising the English version of the article.

Funding

This work was supported by the European Commission, EU-FP7-Systems Microscopy Network of Excellence (grant agreement number 258068) and FP7-INFRASTRUCTURES-BioMedBridges (grant agreement number 284209), the Spanish Ministry of Economy and Competitiveness with European Regional Development Fund (SAF2016-78041-C2-1- R) and the Andalusian Government with European Regional Development Fund (CTS-486). The CIBERER is an initiative from the Instituto de Salud Carlos III.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Authors' contributions

JAGR supervised the project. JAGR and JKH designed the experiments. ADR and JKH gave mathematical advice to the project. BSS performed the experiments and data analysis. JKH wrote the article. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent to publish

Not applicable.

Ethics and consent to participate

Not applicable.

Author details

[1] Department of Molecular Biology and Biochemistry, University of Málaga, Boulevard Louis Pasteur, 29071 Málaga, Spain. [2] Department of Algebra, Geometry and Topology, University of Málaga, Boulevard Louis Pasteur, 29071 Málaga, Spain. [3] European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [4] CIBER de Enfermedades raras (CIBERER), Madrid, Spain.

Received: 27 May 2016 Accepted: 28 January 2017
Published online: 10 February 2017

References

1. Lock JG, Strömblad S. Systems microscopy: an emerging strategy for the life sciences. Exp Cell Res. 2010;316(8):1438–44. doi:10.1016/j.yexcr.2010.04.001.
2. Neumann B, Walter T, Hériché JK, Bulkescher J, Erfle H, Conrad C, Rogers P, Poser I, Held M, Liebel U, Cetin C, Sieckmann F, Pau G, Kabbe R, Wünsche A, Satagopam V, Schmitz MHA, Chapuis C, Gerlich DW, Schneider R, Eils R, Huber W, Peters JM, Hyman AA, Durbin R, Pepperkok R, Ellenberg J. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature. 2010;464(7289):721–7. doi:10.1038/nature08869.
3. Fuchs F, Pau G, Kranz D, Sklyar O, Budjan C, Steinbrink S, Horn T, Pedal A, Huber W, Boutros M. Clustering phenotype populations by

genome-wide RNAi and multiparametric imaging. Mol Syst Biol. 2010;6(370):370. doi:10.1038/msb.2010.25.
4. Simpson JC, Joggerst B, Laketa V, Verissimo F, Cetin C, Erfle H, Bexiga MG, Singan VR, Hériché JK, Neumann B, Mateos A, Blake J, Bechtel S, Benes V, Wiemann S, Ellenberg J, Pepperkok R. Genome-wide RNAi screening identifies human proteins with a regulatory function in the early secretory pathway. Nat Cell Biol. 2012;14(7):764–74. doi:10.1038/ncb2510.
5. Hoehndorf R, Harris MA, Herre H, Rustici G, Gkoutos GV. Semantic integration of physiology phenotypes with an application to the Cellular Phenotype Ontology. Bioinforma (Oxford, England). 2012;28(13):1783–9. doi:10.1093/bioinformatics/bts250.
6. Jupp S, Malone J, Burdett T, Hériché JK, Williams E, Ellenberg J, Parkinson H, Rustici G. The cellular microscopy phenotype ontology. J Biomed Semant. 2016;7(1):28. doi:10.1186/s13326-016-0074-0.
7. Hériché JK, Lees JG, Morilla I, Walter T, Petrova B, Roberti MJ, Hossain MJ, Adler P, Fernandez JM, Krallinger M, Haering CH, Vilo J, Valencia A, Ranea JA, Orengo C, Ellenberg J. Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. Mol Biol Cell. 2014;25(16):2522–36. doi:10.1091/mbc.E13-04-0221.
8. Moudry P, Lukas C, Macurek L, Neumann B, Heriche JK, Pepperkok R, Ellenberg J, Hodny Z, Lukas J, Bartek J. Nucleoporin NUP153 guards genome integrity by promoting nuclear import of 53BP1. Cell Death Differ. 2012;19(5):798–807. doi:10.1038/cdd.2011.150.
9. Balestra F, Strnad P, Flückiger I, Gönczy P. Discovering regulators of centriole biogenesis through siRNA-based functional genomics in human cells. Dev Cell. 2013;25(6):555–71. doi:10.1016/j.devcel.2013.05.016.
10. Paulsen RD, Soni DV, Wollman R, Hahn AT, Yee MC, Guan A, Hesley JA, Miller SC, Cromwell EF, Solow-Cordero DE, Meyer T, Cimprich KA. A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability. Mol Cell. 2009;35(2):228–39. doi:10.1016/j.molcel.2009.06.021.
11. Schmidt EE, Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. Nucleic Acids Res. 2013;41(D1):1021–6. doi:10.1093/nar/gks1170.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9. doi:10.1038/75556.
13. Kirsanova C, Brazma A, Rustici G, Sarkans U. Cellular phenotype database: a repository for systems microscopy data. Bioinformatics. 2015;31(16):2736–40. doi:10.1093/bioinformatics/btv199.
14. Wild F. Lsa: Latent Semantic Analysis. 2015. R package version 0.73.1. https://CRAN.R-project.org/package=lsa.
15. Hennig C, Hausdorf B. Prabclus: functions for clustering of presence-absence, abundance and multilocus genetic data. 2015. R package version 2.2-6. https://CRAN.R-project.org/package=prabclus.
16. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc. 2004;60(5):503–20. doi:10.1108/00220410410560582.
17. Fang H, Gough J. The 'dnet' approach promotes emerging research on cancer patient survival. Genome Med. 2014;6:64. doi:10.1186/s13073-014-0064-8.
18. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Del-Toro N, Duesbury M, Dumousseau M, Galeota E, Hinz U, Iannuccelli M, Jagannathan S, Jimenez R, Khadake J, Lagreid A, Licata L, Lovering RC, Meldal B, Melidoni AN, Milagros M, Peluso D, Perfetto L, Porras P, Raghunath A, Ricard-Blum S, Roechert B, Stutz A, Tognolli M, Van Roey K, Cesareni G, Hermjakob H. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42(D1):358–63. doi:10.1093/nar/gkt1115.
19. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stümpflen V, Mewes HW, Ruepp A, Frishman D. The MIPS mammalian protein-protein interaction database. Bioinformatics. 2005;21(6):832–4. doi:10.1093/bioinformatics/bti115.
20. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 2004;32(Database issue):449–51. doi:10.1093/nar/gkh086.
21. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Database issue):535–9. doi:10.1093/nar/gkj109.
22. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, Stein L. Annotating cancer variants and anti-cancer therapeutics in Reactome. Cancers. 2012;4(4):1180–211. doi:10.3390/cancers4041180.
23. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2016;44(D1):481–7. doi:10.1093/nar/gkv1351.
24. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. Nucleic Acids Res. 2014;42(D1):396–400. doi:10.1093/nar/gkt1079.
25. Trabuco LG, Betts MJ, Russell RB. Negative protein-protein interaction datasets derived from large-scale two-hybrid experiments. Methods. 2012;58(4):343–8. doi:10.1016/j.ymeth.2012.07.028.
26. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. Proc: an open-source package for r and s+ to analyze and compare roc curves. BMC Bioinforma. 2011;12:77.
27. Razick S, Magklaras G, Donaldson IM. iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinforma. 2008;9:405. doi:10.1186/1471-2105-9-405.
28. Glass K, Girvan M. Finding New Order in Biological Functions from the Network Structure of Gene Annotations. PLoS Comput Biol. 2015;11(11): 1004565. doi:10.1371/journal.pcbi.1004565.
29. Meila M, Shi J. A Random Walks View of Spectral Segmentation. In: Proceedings of the International Conference on AI and Statistics (AISTATS); 2001. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.30.8065.
30. Laufer C, Fischer B, Billmann M, Huber W, Boutros M. Mapping genetic interactions in human cancer cells with RNAi and multiparametric phenotyping. Nat Methods. 2013;10(5):427–31. doi:10.1038/nmeth.2436.
31. Bakal C, Church G, Perrimon N. Regulating Cell Morphology. Science. 2007;316(June):1753–6.
32. Loo LH, Wu LF, Altschuler SJ. Image-based multivariate profiling of drug responses from single cells. Nat Methods. 2007;4(5):445–53. doi:10.1038/nmeth1032.
33. Wang X, Castro MA, Mulder KW, Markowetz F. Posterior association networks and functional modules inferred from rich phenotypes of gene perturbations. PLoS Comput Biol. 2012;8(6):1–16. doi:10.1371/journal.pcbi.1002566.
34. Gunsalus KC, Yueh WC, MacMenamin P, Piano F. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. Nucleic Acids Res. 2004;32(Database issue):406–10. doi:10.1093/nar/gkh110.
35. Groth P, Weiss B, Pohlenz HD, Leser U. Mining phenotypes for gene function prediction. BMC Bioinforma. 2008;9(1):136. doi:10.1186/1471-2105-9-136.
36. Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. Mol Syst Biol. 2007;3(88):88. doi:10.1038/msb4100129.
37. Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: Assessment with biological features and issues. Brief Bioinform. 2012;13(5):569–85. doi:10.1093/bib/bbr066.
38. Lees JG, Hériché JK, Morilla I, Fernandez JM, Adler P, Krallinger M, Vilo J, Valencia A, Ellenberg J, Ranea JA, Orengo C. FUN-L: gene prioritization for RNAi screens. Bioinformatics. 2015;31(12):2052–3.
39. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome Biol. 2008;9 Suppl 1:4. doi:10.1186/gb-2008-9-s1-s4.
40. Fischer B, Sandmann T, Horn T, Billmann M, Chaudhary V, Huber W, Boutros M. A map of directional genetic interactions in a metazoan cell. eLife. 2015;2015(4):1–21. doi:10.7554/eLife.05464.
41. Billmann M, Horn T, Fischer B, Sandmann T, Huber W, Boutros M. A genetic interaction map of cell cycle regulators. Mol Biol Cell. 2016;27(8): 1397–407. doi:10.1091/mbc.E15-07-0467.

Serrano-Solano et al. BMC Bioinformatics
DOI 10.1186/s12859-017-1607-y

BMC Bioinformatics

ERRATUM Open Access

Erratum to: How can functional annotations be derived from profiles of phenotypic annotations?

Beatriz Serrano-Solano[1], Antonio Díaz Ramos[2], Jean-Karim Hériché[3] and Juan A. G. Ranea[1,4*]

Erratum

Upon publication of this article [1], it was brought to our attention that Table 1 was incorrectly presented. The correct Table 1 is shown below and has been updated in the original article.

Author details
[1]Department of Molecular Biology and Biochemistry, University of Málaga, Boulevard Louis Pasteur, 29071 Málaga, Spain. [2]Department of Algebra, Geometry and Topology, University of Málaga, Boulevard Louis Pasteur, 29071 Málaga, Spain. [3]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [4]CIBER de Enfermedades raras (CIBERER), Madrid, Spain.

Received: 15 March 2017 Accepted: 15 March 2017
Published online: 27 March 2017

Reference

1. Serrano-Solano B, et al. How can functional annotations be derived from profiles of phenotypic annotations? BMC Bioinformatics. 2017;18:96. doi:10.1186/s12859-017-1503-5.

* Correspondence: ranea@uma.es
[1]Department of Molecular Biology and Biochemistry, University of Málaga, Boulevard Louis Pasteur, 29071 Málaga, Spain
[4]CIBER de Enfermedades raras (CIBERER), Madrid, Spain

Table 1 Set of 36 phenotypes obtained from the listed siRNA experiments sorted by its CMPO identifier

Experiment	Description	Phenotypes	IDs in CMPO
CellMorph [3]	Genome-wide RNAi screen that examines changes in the morphology of individual HeLa cells within cell populations.	- Decreased cell number	CMPO:0000052
		- Cell with projections	CMPO:0000071
		- Elongated cell	CMPO:0000077
		- More lamellipodia cells	CMPO:0000083
		- Increased number of actin filament	CMPO:0000105
		- Round cell	CMPO:0000118
		- Increased cell size	CMPO:0000128
		- Decreased cell size	CMPO:0000129
		- Bright nuclei	CMPO:0000154
		- Metaphase arrested	CMPO:0000305
		- Increased cell size in population	CMPO:0000340
MitoCheck [2]	Genome-wide RNAi screen for genes required for chromosome segregation in HeLa cells. The screen also reports genes involved in other processes such as cell movement.	- Cell death	CMPO:0000030
		- Increased nucleus size	CMPO:0000140
		- Graped micronucleus	CMPO:0000156
		- Abnormal nucleus shape	CMPO:0000157
		- Mitosis delayed	CMPO:0000202
		- Binuclear cell	CMPO:0000213
		- Absence of mitotic chromosome decondensation	CMPO:0000216
		- Increased cell movement speed	CMPO:0000236
		- Increased cell movement distance	CMPO:0000237
		- Proliferating cells	CMPO:0000241
		- Metaphase delayed	CMPO:0000307
		- Abnormal chromosome segregation	CMPO:0000326
		- Prometaphase delayed	CMPO:0000344
		- Increased variability of nuclear shape in population	CMPO:0000345
		- Mitotic metaphase plate congression	CMPO:0000348
EMBL secretion [4]	Genome-wide RNAi screen for interference with ER-to-plasma membrane transport of the secretory cargo protein tsO45G in HeLa cells.	- Increased rate of protein secretion	CMPO:0000246
		- Mild decrease in rate of protein secretion	CMPO:0000318
		- Strong decrease in rate of protein secretion	CMPO:0000319
		- Decreased rate of intracellular protein transport	CMPO:0000346
GR00053 [10]	Genome-wide RNAi screen for genes involved in DNA damage responses in HeLa cells.	- Increased number of site of double-strand break	CMPO:0000182
GR00290 [9]	Genome-wide RNAi screen for genes regulating centriole formation in HeLa cells.	- Increased centriole replication	CMPO:0000361
		- Decreased centriole replication	CMPO:0000362
Copenhagen DNA damage Ubiquitin [8]	RNAi screen of >1300 genes involved in the ubiquitin-proteasome system or encoding zinc-finger proteins looking for modulators of cellular responses to ionizing radiation in HeLa and U2OS cells.	- Decreased number of site of double-strand brea	CMPO:0000181
EMBL chromosome condensation [7]	RNAi screen of 100 bioinformatically-selected genes for changes in mitotic prophase duration in HeLa cells.	- Increased duration of mitotic prophase	CMPO:0000328
		- Decreased duration of mitotic prophase	CMPO:0000329